The Dissertation Committee for Alexander Timothy Hawk certifies that this is the approved version of the following dissertation:

Calculating rare biophysical events. A study of the milestoning method and simple polymer models.

Committee:

Dmitrii Makarov, Co-Supervisor

Ernst-Ludwig Florin, Co-Supervisor

Ron Elber

George Shubeita

Charles Chiu

Graeme Henkelman

Calculating rare biophysical events. A study of the milestoning method and simple polymer models.

by

Alexander Timothy Hawk B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2012

Dedicated to the two people who are most responsible for my success and happiness, my parents Michal and Timothy Hawk

Acknowledgements

I would first like to acknowledge my PhD supervisor Dmitrii Makarov, for giving me an unquantifiable amount of scientific knowledge, timely feedback on papers, motivation, and support. His passion for science is inspirational, and I am a better man having worked with him. I would also like to acknowledge my co-supervisor, Ernst-Ludwig Florin for helping me to become a better speaker and always giving me sound and true advice. His group parties, featuring jam sessions and sensational grilling (a special shout out to "Big Fella"), were among the best I have been to here in Austin. While Ron Elber was not officially a supervisor, I am also grateful to him for pioneering one of the most promising techniques in the molecular dynamics community (Milestoning), and providing me with invaluable feedback and insight throughout my graduate studies. I am also grateful to my former supervisor, Charles Chiu, who motivated me by his own enthusiasm to do "frontier research". I am thankful to my undergraduate advisors, Yitzhak Tor and Stephen Leone for introducing me to the world of science and cultivating my sensibilities at a critical period in my professional development. I can still hear their words of advice that have been proven oh so true over the past few years.

I don't think I could have had better officemates than I did in Sai Konda and Ryan Cheng. Much of what I learned in grad school came in conversations (often unexpected) from these talented and entertaining gentleman. Former Makarov Group members Serdal Kirmizialtin and Lei Huang were also both a pleasure knowing and extremely helpful through out my graduate studies. Similarly, I had fantastic group mates in Dr. Florin's group: Vassili Demergis, Andrea Keidel, Chieze Ibeneche, Martin Kochanczyk, Tobias

Bartsch, Katie Hinko, Katja Taute, and Pinyu Thrasher, who helped me prepare for talks and were always a pleasure to join for happy hours.

And last but certainly not least I'd like to thank my family and friends. I thank my parents who raised me to trust the direction of my heart and to value hard work; and I'm thankful to my amazing sisters for their confidence and making the effort to stay close while I have been away from home. Finally, for brightening and enriching my life, I'd like to acknowledge my friends back home: Bob Sise and David Hicks, my friends here in Texas: Sergey Maslennikov, Omar Ahmed, Dong Hyun Kim, Denys Maslov, Hannah Bowman, James Murdock, Ryan Cheng, Sai Konda, Saul Jerome San Juan, Alan Robinson, and my friends in the Austin Sacred Harp singers group lead by Gaylon Powel and Leon Ballinger.

Calculating rare biophysical events. A study of the milestoning method and simple polymer models.

Alexander Timothy Hawk, Ph.D.

The University of Texas at Austin, 2012

Supervisors: Dmitrii Makarov and Ernst-Ludwig Florin

Performing simulations of large-scale bio-molecular systems has long been one of the great challenges of molecular biophysics. Phenomena, such as the folding and conformational rearrangement of proteins, often takes place over the course milliseconds-to-seconds. The methods of traditional molecular dynamics used to simulate such systems are on the other hand typically limited to giving trajectories of nanosecond-to-microsecond duration. The failure of traditional methods has thus motivated the development of many special purpose techniques that propose to capture the essential characteristics of systems over conventionally inaccessible timescales.

This dissertation first focuses on presenting a set of advances made on one such technique, Milestoning. Milestoning gives a statistical procedure for recovering long trajectories of the system based on observations of many short trajectories that start and end on hypersurfaces in the system's phase space. Justification of the method's validity typically relies on the assumption that trajectories of the system lose all memory between

crossing successive milestones. We start by giving a modified milestoning procedure in which both the memory loss assumption is relaxed and reaction mechanisms are more easily extracted. We follow with numerical examples illustrating the success of new procedure. Then we show how milestoning may be used to compute an experimentally relevant timescale known as the transit time (also known as the reaction path time). Finally, we discuss how time reversal symmetry may be exploited to improve sampling of the trajectory fragments that connect milestones.

After discussing milestoning, the dissertation shifts focus to a different way of approaching the problem of simulating long timescales. We consider two polymers models that are sufficiently simple to permit numerical integration of the desired long trajectories of the system. In some limiting cases, we see their simplicity even permits some questions about the dynamcis to be answered analytically. Using these models, we make a series of experimentally verifiable predictions about the dynamics of unfolded polymers.

Table of Contents

List of Figures

Chapter 1. Introduction

In computational chemistry, molecular simulations are used to reveal system properties that cannot be easily measured in experiments. Reaction pathways, rate coefficients, and free energy landscapes are just a few examples of things that may be computed via standard techniques. Currently, one of the greatest challenges in the field is achieving simulations of large-scale physical systems over experimentally relevant timescales. In molecular biophysics, this problem is of particular importance because phenomena such as protein folding[1] and RNA folding[2], may take place at millisecond-to-second timescales while typical molecular dynamics (MD) simulations conducted with modern computers are typically limited to producing nano-to-microsecond trajectories. With massive parallelization and special purpose machines, millisecond trajectories can be generated[3]. Still, even with constantly increasing computational resources, collecting enough of these trajectories to adequately sample the various reaction pathways remains extremely challenging. This dissertation concentrates on approaches to obtaining such long trajectories in the scope of classical molecular dynamics.

In classical molecular dynamics, physical systems are represented by point particles that evolve according to Newton's laws. In the most detailed simulations, each atom in the system is represented as a particle that interacts with the other particles through a potential energy function that accounts for various relevant interactions—e.g. electrostatics, van der Waals, the interactions between bonded pairs of atoms, etc. For

sufficiently large systems (e.g. proteins) however, the computational costs of such a detailed description become prohibitively expensive, limiting trajectories to the nano-to-microsecond timescale.

Many alternative techniques to brute force molecular dynamics have been proposed to account for the behavior of molecular systems on long timescales. Of these, there are essentially two categories: Those which focus on exploiting properties of the system's energy landscape to simplify calculations of particular kinetic quantities (e.g. the rate constant of a reaction), and those which focus on representing the system with fewer degrees of freedom (e.g. groups of atoms might be represented by a single particle in the simulation). In the second and third chapters, we present advances made on a technique, Milestoning, that is of the first category; and in the fourth chapter, we present a set of experimentally verifiable predictions made from consideration of simple polymer models, which exclude explicit representation of the vast majority of degrees of freedom present in typical polymers.

1.1 Milestoning

Milestoning is a method that reconstructs long trajectories of the system based on observations of short trajectories that start and end on hypersurfaces in the system's phase space. Typical application of the method is justified by an assumption that trajectories lose memory[4-6] in the time between crossing successive milestones, which loosely put means that the coordinates a trajectory has at the time of arrival on a

milestone do not affect the probabilities of transition to the various other milestones. When the energy landscape of the system is not particularly jagged (e.g. the dynamics are diffusive) the assumption is often justified for suitably defined milestones. Indeed milestoning has been shown to be quite accurate for a variety of diffusive processes[4, 7-9]. With regard to milestoning, we present three findings. First, we show that with slight modification of the milestoning procedure, it may be applied to processes that do not well satisfy the conditions that justify application of the original method. Furthermore, we also show the modified method may be used to improve the method's capacity for extracting reaction mechanism. Second, we show how milestoning may be used to compute a timescale known as the transit time, which is a measure of the time successful reactive trajectories spend in transition between the reactants and products. As the transit time has been the subject of many recent experimental studies[10-14], the milestoning algorithm is poised to permit comparison of simulation and experiment where not previously possible. Finally, we introduce a procedure that exploits time reversal symmetry to extract more information from the molecular dynamics simulations performed in milestoning. In principle the technique may be used to reduce the number of required MD simulations in a given milestoning calculation.

1.2 Polymer models with internal friction

The simple polymer models discussed in the fourth chapter are the Rouse with Internal Friction (RIF) and Zimm with internal friction (ZIF) models. They describe the

motion of an unfolded polymer in a fluid of low Reynolds number (e.g. dynamics are overdamped), with the difference being that the ZIF model accounts for hydrodynamic interactions between the monomeric units of the polymer chain while the RIF model does not. While the RIF model has been studied in detail previously[15-19], treatment of the ZIF model has been limited to qualitative discussion[18]. Motivated by a dearth of adequate theoretical models that can be used to interpret experimental observations of the unfolded state dynamics of proteins, we develop the ZIF model in detail and then employ both models to make a set of experimentally verifiable predictions.

With our predictions, we aim to clarify the effect internal friction has on the dynamics of unfolded biopolymers. Internal friction, as the name suggests, is a frictional force that arises from a polymer's intrinsic resistance to conformational change. Treating internal friction as a free parameter, we investigate it's effect on two timescales: the timescale of reconfiguration—which is the timescale on which monomer-to-monomer position vectors decorrelate—and the timescale of loop formation—which is the timescale on which the end monomors come into contact. We find that while reconfiguration timescales have a simple linear dependence on internal friction, loop formation timescales exhibit a highly nontrivial dependence. However, we show in the limit of high and low internal friction, loop formation times may be explained with an analytic model and a scaling relationship, respectively.

Chapter 2. Milestoning with Coarse Memory

2.1 Abstract

Milestoning is a method used to calculate the kinetics of molecular processes occurring on timescales inaccessible to traditional molecular dynamics (MD) simulations. In the method, the phase space of the system is partitioned by milestones (hyper-surfaces), trajectories are initialized on each milestone, and short MD simulations are performed to calculate transitions between neighboring milestones. Long trajectories of the system are then reconstructed with a semi-Markov process from the observed statistics of transition. The procedure is typically justified by the assumption that trajectories lose memory between crossing successive milestones. Here we present Milestoning with coarse memory (MCM), a generalization of milestoning that relaxes the memory loss assumption of conventional milestoning. In the method, milestones are defined and sample transitions are calculated in the standard milestoning way. Then, milestones are broken up into distinct neigborhoods (clusters), and each sample transition is associated with two clusters: the cluster containing the coordinates the trajectory was initialized in, and the cluster (on the terminal milestone) containing trajectory's final coordinates. Long trajectories of the system are then reconstructed with a semi-Markov process in an extended state space built from milestone and cluster indices. To test the method, we apply it to a process that is particularly ill suited for milestoning: the dynamics of a polymer confined to a narrow cylinder. We show that milestoning calculations of both the mean first passage time (MFPT) and the mean transit time (MTT)

of polymer reversal, which occurs when the end-to-end vector reverses direction, are significantly improved when MCM is applied, and further, information about the mechanism of polymer reversal may be extracted using MCM. Also, we see the overhead of performing MCM on top of conventional milestoning is negligible.

2.2 Introduction

Milestoning[4, 5, 20-23] is a method that has been used to approximate the kinetics of a variety processes in molecular biophysics, such as the dynamics of molecular motors[24], enzyme substrate association[9], and protein unfolding[6], helix formation[7], protein permeation through a membrane[8], which take place on microsecond-to-hour timescales—well beyond the limits of traditional molecular dynamics (MD) simulations. Among many methods[4, 5, 20, 22, 23, 25-35] that propose to extend the timescales of molecular simulations, milestoning stands out as both exceedingly efficient, and robust. Unlike methods such as Transition Path Sampling[25], Transition Interface Sampling[30], and Partial Path Transition Interface Sampling[26], milestoning neither requires the system to be governed by exponential kinetics nor be globally equilibrated—though conventional milestoning does assume the system relaxes to local equilibrium in various regions of phase space referred to as milestones. Forward Flux is a method that poses no equilibration requirements on the system, but suffers from the assumption of exponential kinetics and furthermore requires constructing full length trajectories of the long timescale process of interest. Milestoning is most similar to the Markov State Model

(MSM)[31-35], as both give statistical procedures for recovering long trajectories of the system based on observations of short trajectory fragments. However, there are differences in how the methods compute trajectory fragments, and differences in the algorithms used to extract the long time kinetics of the system.

In milestoning, the phase space of an *N* particle system is partitioned by milestones (hypersurfaces in the system's phase space) and the *state* of a trajectory is given by the last milestone visited. The assumption one typically makes to apply milestoning is that trajectories lose memory in the time between transitions—which put loosely means that the particular coordinates a trajectory assumes when it arrives at a milestone do not affect the probabilities that govern it's future transitions. When this assumption is satisfied, the transition probabilities between milestones and the probability distribution of lag times between transitions are simple quantities that can be plugged in to standard algorithms[5, 20, 21] that yield many of the relevant kinetic quantities of the system, such as the mean first passage time (MFPT) and the mean transit time (MTT), which is the mean time for the system to pass from the reactants to the products, assuming it never returns to the reactants after entering the transition region between the reactants and the products. Fortunately, the transition probabilities and lag time distributions are not themselves difficult to compute, even for milestones that lie in regions of low probability, as the sample transitions used to compute these quantities are given by trajectories that are initialized directly on the milestones and followed for only a short period of time until they arrive at different milestones.

While the memory loss assumption is sufficient to ensure the accuracy of milestoning calculations, we note that it is not necessary. For example, if one is only interested in obtaining the MFPT of a process in equilibrium, then it is only required that the sequence of milestones visited by a long trajectory is given by a discrete Markov chain[21]. For ergodic systems undergoing Brownian dynamics, the MFPT is thus given exactly when milestones are taken as iso-commitor surfaces—surfaces where a trajectory's probabilities of going forward or backwards to the neighboring milestones are independent of it's coordinates on the milestone—regardless of whether trajectories lose memory between successive milestones. However, finding iso-committor surfaces in practice is not always computationally feasible, especially for high dimensional systems. Therefore in practice, milestones are typically chosen such that the memory loss assumption is reasonable.

As discussed in a previous study by West et al.[5], the computational gain realized in milestoning calculations depends heavily on how milestones are spaced in the system's phase space, where closer spacing leads to shorter MD simulations and greater efficiency. At the same time however, trajectories are less likely to satisfy the memory loss assumption the closer milestones are to one another. To remedy this conflict, we present a generalization of milestoning: Milestoning with Coarse Memory (MCM). In this method, the phase space of a milestone is broken into C neighborhoods referred to as clusters, and the state of a trajectory is then given by a milestone, cluster pair: (i,c), where c is the index of the cluster that the trajectory first sampled when it arrived at

milestone *i* from some other milestone. By describing a trajectory's state in greater detail, we are able to justify application of MCM with a significantly weaker assumption of memory loss than what is made in conventional milestoning. We remark that MCM is not the first method to relax the memory assumption, as the M1TM method previously has offered a procedure for doing so. However, we show in the present manuscript that M1TM is but a special case of MCM.

Another motivation for developing MCM is that the method enables one to analyze the behavior of trajectories in greater detail, as clusters may be associated with characteristic conformations and ranges of values for variables relevant to the reaction. Already, much mechanistic information may be extracted from directional milestoning (DiM) calculations, as the points on each milestone tend to be "close" (according to some metric on the configuration space) to a particular conformation or class of conformations, referred to as an anchor. The limitation one faces in a DiM calculation though is that the anchors must be specified before performing MD simulations. If it turns out the reactions mechanism cannot be discerned from the dynamics between the chosen anchors, the full calculation must be repeated (MD simulations included) with new anchors. In MCM, the clusters, like directional milestones, may be defined such that the points in a cluster are close to anchors (though we note there are ways of defining clusters without anchors). If the initial choice of anchors in an MCM calculation does not permit the extraction of a reaction mechanism, the clusters may be redefined in terms of different anchors as many

times as is desired so to permit a comprehensive investigation of the reaction mechanism, without requiring additional MD simulations.

We organize the rest of manuscript as follows. In section 2.3 we develop the theory of MCM and also discuss a variety of approaches to implementation. In section 2.4, we apply the method to the dynamics of a polymer confined to a long cylindrical pore, a process in which position memory does not fully decay in the time period between transitions. We demonstrate that even with a relatively simple clustering procedure, the accuracy of milestoning calculations of the MFPT and the MTT of polymer reversal[36, 37] (the event in which the polymer reverses its end-to-end vector) may be significantly improved by applying MCM. Section 2.5 contains some concluding remarks, and 2.6 contains the chapter Appendix.

2.3 Milestoning with Coarse Memory

2.3.1 Theory

In milestoning, M regions of phase space (typically hyper-surfaces) called milestones, $H_1, H_2, ..., H_M$, partition the phase space of the molecular system. The state of a trajectory is given by the integer index, $i(t)$, of the milestone it last visited, and its evolution is specified by the sequence of milestones visited, $\{i_1, i_2, ..., i_n, ...\}$, and the sequence of times spent in each state (lag times), $\{s_1, s_2, ..., s_n, ...\}$. In MCM, like in conventional milestoning, the trajectory is said to make a transition each time it samples a

milestone that is different than the last. However, the state of the trajectory is no longer specified by $i(t)$ alone. Rather, it is specified by an ordered pair (i,c), where *i* represents the last milestone the trajectory sampled, and *c* labels the "cluster" (the subregion of milestone *i*) that the trajectory sampled when it made the transition to *i* (Fig. 2.1). As seen in Fig 2.1, the trajectory then remains in a given MCM state until it samples a different milestone, even if the trajectory goes onto sample a different cluster on milestone *i*. The reasoning behind our rule for transitions is given in the next subsection, and procedures for defining clusters will be given in subsections thereafter.

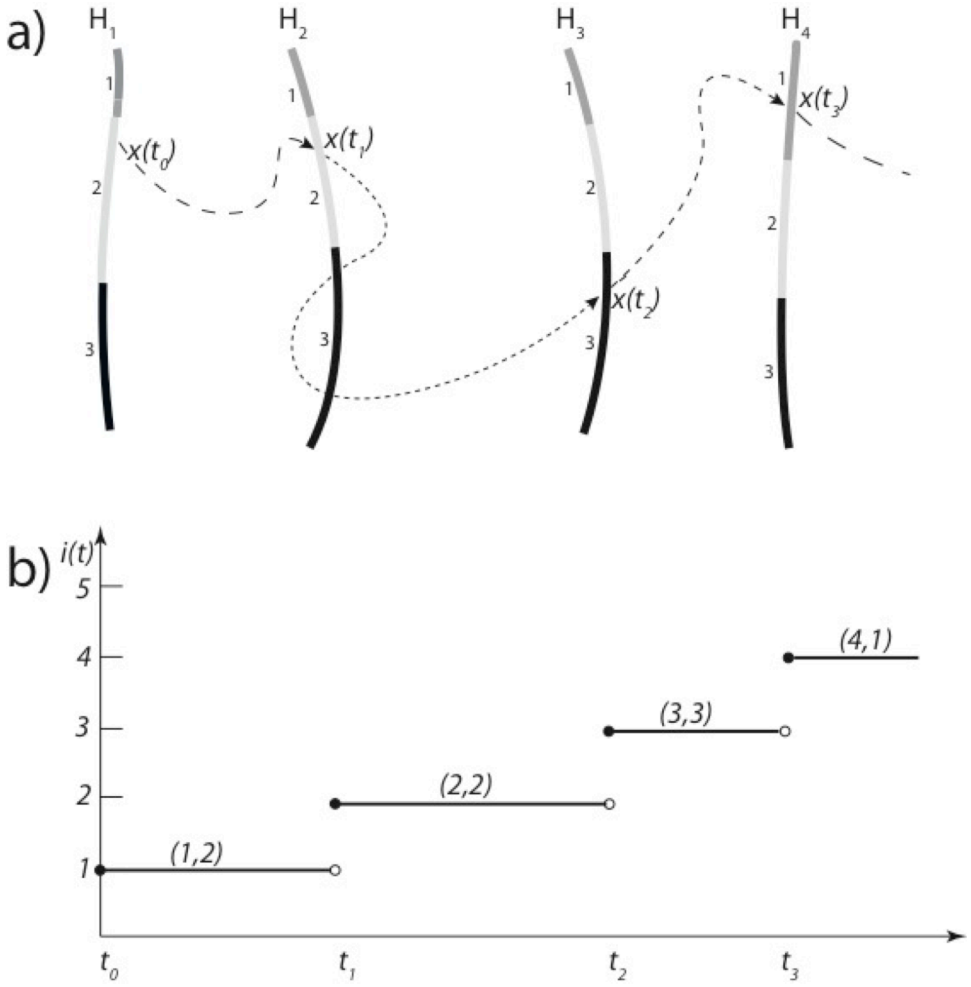


Figure 2.1: 1a is a schematic representation of a trajectory $\mathbf{x}(t)$ that crosses milestones H_1, H_2, H_3, H_4. Each milestone is divided into 3 numbered regions (clusters), distinguished by their different opacity. In 1b, the piecewise continuous function $i(t)$ tracks the index of the milestone the trajectory in 1a last crossed, and its value is the first index in the ordered pair above each continuous portion of $i(t)$. The 2nd index of each ordered pair denotes the cluster the trajectory sampled when it transitioned to the current milestone. These ordered pairs define the MCM state of the molecular system.

For now, we focus on the assumptions that make MCM a rigorously valid description of the system dynamics, noting they are fundamentally the same as those justifying milestoning. Letting $\alpha(t)$ denote the state of the system at time *t*, both MCM and conventional milestoning require that $\alpha(t)$ for a long trajectory of the system is given by a semi-Markov process. That is, to say they require that:

(C1) The sequence of states, $\{\alpha_1, \alpha_2, ..., \alpha_n, ...\}$, is given by a discrete Markov chain.

(C2) The lag time, s_n, between two successive transitions is a random number, whose probability distribution $\pi_{\alpha_{n+1},\alpha_n}(s_n)$ depends only on the current state α_n, and next state α_{n+1}.

When these conditions are satisfied, the evolution of the system may be found through the coupled set of integro-differential equations[5, 20]

$$p_\alpha(t) = \int_0^t Q_\alpha(t') \left[1 - \int_0^{t-t'} \sum_{\beta \in \bar{\alpha}} K_{\beta\alpha}(t'') dt'' \right] dt' \tag{2.1}$$

$$Q_\beta(t) = p_\beta(0)\delta(t^+) + \sum_{\alpha \in \bar{\beta}} \int_0^t Q_\alpha(t') K_{\alpha\beta}(t - t') dt' \tag{2.2}$$

$$K_{\alpha\beta}(t) \equiv T_{\beta\alpha} \pi_{\beta\alpha}(t), \tag{2.3}$$

where $T_{\beta\alpha}$ is defined as the conditional probability for the system to make a transition from α to β, and $\bar{\beta}$ represents the set of states from which a trajectory may reach β by making a single transition. We remark that the sub-indices are transposed from the left

hand side to the right hand side of Eq. 2.3 to be consistent with conventions for these symbols in the literature[9, 23]. Furthermore, efficient procedures for obtaining the MFPT and the MTT, which do not require integration of Eqs. 2.1 and 2.2, have been given in the literature[4, 5, 20, 21, 38].

A general approach to satisfying C1 and C2 is to choose states such that memory is lost between transitions. To give this condition more precisely, we first define the transition kernel $K_{\alpha\beta}\left(x_\alpha, x_\beta; t\right)$ as the conditional probability density that a trajectory initialized at time 0 at the point x_α (in phase space) in state α will transition directly to state β, arriving there at the point x_β at time *t*. We say that the memory loss condition is satisfied if:

$$K_{\alpha\beta}\left(x_\alpha, x_\beta; t\right) \simeq K_{\alpha\beta}\left(x_\beta; t\right). \tag{2.4}$$

Indeed, as shown in Appendix 2.5.1, Eq. 2.4 is sufficient to ensure that Eqs. 2.1 and 2.2 accurately describe the evolution of a long trajectory of the system in the state space. Generic as Eq. 2.4 seems, it applies quite differently however to MCM and conventional milestoning. Applied to MCM, it poses the requirement that trajectories starting at different phase space points in the same cluster have identical properties of transition, where as with conventional milestoning, it requires that trajectories starting at different points anywhere on the milestone have the same properties of transition.

With the fundamental equations of MCM now established, we turn our attention to the nuts and bolts of how transitions between MCM states are sampled, and after that to how clusters on milestones may be defined.

2.3.2 Initial configurations and transitions

We start by reviewing the standard milestoning procedure for obtaining an ensemble of initial configurations (which are phase space points on a milestone, in the context of this work) as this procedure is a necessary part of MCM. In milestoning, the initial configurations on a milestone are sampled from it's first hitting point distribution, which is the probability distribution over the points in phase space a trajectory may sample upon transition to the milestone. A variety of procedures have been given for sampling this distribution[4, 23]. In each, configurations are sampled on the milestone under equilibrium conditions; trajectories are initialized in these configurations and followed backwards in time to the previous milestone crossed. If this previous milestone is different than the milestone the trajectory was initialized on, the trajectory's initial configuration is indeed a first hitting point, and it is added to the ensemble of initial configurations on the milestone.

Transitions are then sampled by following trajectories forward in time from each initial configuration until they arrive at a new milestone. In standard milestoning, the transition probabilities and lag time distributions are immediately computed from the statistics afforded by the sample transitions. In MCM however, we add a post processing stage where in which we associate each trajectory with MCM states on the initial and

final milestones. First, we define clusters on the milestones based on one or more of the techniques discussed in section 2.2.3. Then, for each trajectory that exhibits a transition form milestone *i* to *j*, we identify the cluster, *c,* that contains it's initial configuration on milestone *i*, and the cluster, *d*, that contains it's terminal configuration on milestone *j*. The trajectory is then viewed as a sample transition between the MCM states (i,c) and (j,d). With each sample yielding a well defined transition in the MCM state space, the MCM transition probabilities and lag time distributions may be computed. Provided the computed transition probability matrix is such that the ergodic theorem[39] is satisfied (i.e. the state space does not have two or more subsets which are dynamically disconnected), Eqs. 2.1 and 2.2 may be solved for the evolution of the system.

We remark here that trajectories in MCM do not make transitions when they go between clusters on the same milestone. Consideration of such transitions is unnecessary because as discussed above, each trajectory sampled in the standard milestoning already yields a sample transition between MCM states. Furthermore, allowing transitions between clusters on the same milestone is problematic because there is no guarantee that trajectories will have sufficient time and space over which to lose memory between transitions. In the clustering methods discussed in the next section (e.g. centroid clustering, histogram clustering) clusters are defined in such a way where the distance (in phase space) between points in neighboring clusters on the same milestone may be arbitrarily small.

2.3.3 Clustering

In MCM we define the clusters on the milestones only after all MD data has been collected (i.e. after transitions have been sampled). While it is certainly possible to define the clusters before collecting MD data, doing so is both unnecessary and problematic as it is possible that such predefined clusters may not be visited by the sample trajectories. By waiting till after it is known what regions of the milestones trajectories visit, we have the have the flexibility to define a set of clusters which are all well sampled (e.g. each cluster is sampled by some predefined minimum number of trajectories).

Of the many clustering methods that could conceivably be used, we focus on just three in the present manuscript: centroid, histogram, and transition memory clustering. While the numerical examples in the present manuscript illustrate the latter two, we still discuss centroid clustering as it can in principle be used to satisfy the memory loss assumption in MCM arbitrarily well.

2.3.3.1 Centroid clustering

Centroid clustering[40], also known as k-means or k-centers clustering, is a procedure in which a set of points is partitioned into *k* disjoint sets $\{S_1, S_2, ...S_k\}$. A point *x*, is a member of S_j if the distance between it and the centroid of S_j, a pre-selected point x_j^C, is less than the distance between it and the centroid of any other cluster. Specifically,

$$S_j = \left\{ x \;\middle|\; d\left(x, x_j^C\right) < d\left(x, x_l^C\right) \quad \forall l \right\} \; . \tag{2.5}$$

In application to MCM, the set of points we wish to partition are the initial configurations $\{x_i\}$ on milestone *i*, and the distance function, $d(x,y)$ may be any metric that measures proximity between the configurations *x*, and *y* (e.g. the Euclidean distance between points in phase space). A common choice of centroids is one in which the within-cluster-sum of squares: $\sum_{j=1}^{k} \sum_{x \in S_j} d(x, x_j)$ is minimized, as a number of methods, such as Loyd's algorithm[41], may be used to efficiently search for centroids that satisfy this condition. This

Provided that the region of phase space accessible to the system may be assumed finite on the timescale of process, the distance in phase space between points within clusters may be made arbitrarily small by increasing the number centroids. This suggests that as more clusters are used, the MCM assumption of memory loss may be satisfied arbitrarily well, potentially making MCM an exact description of the system's dynamics in this limit.

One problem with taking the distance function to be a metric on the full phase space of the system, such as the Euclidean distance between configurations, is that coordinates which do not influence the properties of transition (e.g. velocities that quickly de-correlate), may put a large distance between points on the milestone that have similar properties of transition. As a consequence, the number of clusters needed to resolve

states with different properties of transition may simply be too large given sampling limitations.

A better strategy may therefore be to take the distance function as a metric on a few collective variables which decorrelate on a timescale comparable to the characteristic time of transition. If for example $u_1(x)$ and $u_2(x)$ are examples of such variables, an appropriate metric may be $d(x,y) \equiv \sqrt{\left(u_1(x)-u_1(y)\right)^2+\left(u_2(x)-u_2(y)\right)^2}$.

Finally we mention that one problem with using centroid clustering in MCM is that it may give clusters which are sparsely sampled. Fortunately, a variation of it, known as K-Means Clustering with Constraints[42] can be used to define the clusters of the milestones such that they are all equally well sampled by the collected MD data.

2.3.3.2 Histogram clustering

Histogram clustering is a somewhat cruder, but simpler alternative to centroid clustering. In the method, the range of some collective variable, $u(x)$, is partitioned into *k* intervals: $(u_1,u_2]$, $(u_2,u_3]$,...., (u_k,u_{k+1}) that serve to define clusters on a milestone. The interval a trajectory $x(t)$ samples at the time of transition to this milestone determines it's MCM state. We remark that there is no need to select the same variable $u(x)$ for each milestone, and that coordinates that fail to lose memory on the timescale of a typical transition between milestones are natural candidates for $u(x)$.

To ensure good statistics are collected for each interval, it is reasonable to require that an equal number of the sample initial configurations lie in each interval. Intervals can be defined that satisfy this condition by binning the set of initial configurations on the milestone with a fixed height histogram (Fig. 2.2) along *u*.

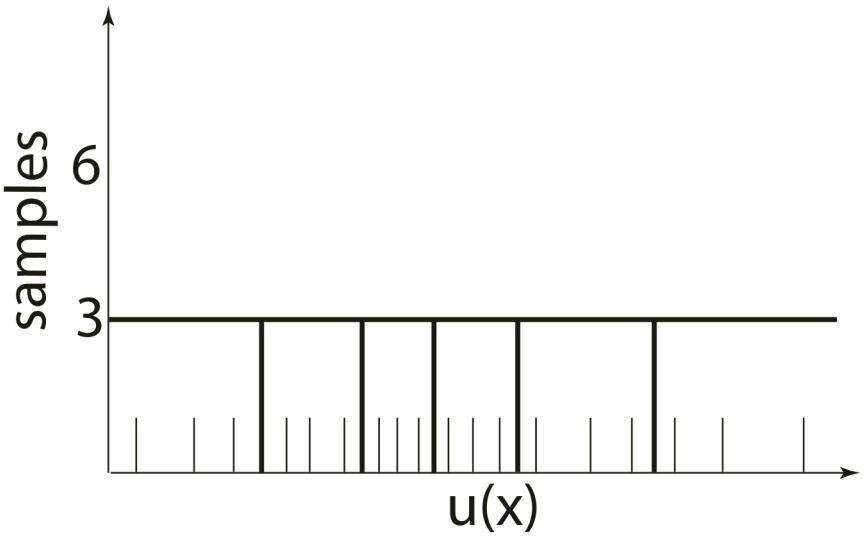


Figure 2.2: A fixed height histogram in which the set of initial configurations on milestone *i*, $\{x_i\}$, is binned along $u(x)$. Each short line is a value of $u(x)$ assumed by a sample initial configuration, and the height of each bin is fixed at 3 samples. The intervals between the thick lines define the clusters on the milestone.

2.3.3.3 Transition memory clustering

Arguably the simplest type of clustering we discuss here is transition memory clustering, which was first used in the M1TM method[23]. In it, the trajectories on milestone *i* are grouped into *k* clusters, where *k* is the number of milestones that are *connected* to *i*. A milestone *j* is said to be connected to *i* if one or more sample trajectories are observed to make transitions from *j* to *i*. Each milestone connected to *i* enumerates an MCM (or M1TM) state on *i*, where trajectories arriving at *i* from *j* belong to the state (i, j).

One nice thing about the method is that it may easily be combined with the previously discussed clustering methods. Doing so just amounts to describing trajectories with slightly more complicated MCM states. Specifically, when combining methods, the MCM states would be enumerated by the set of ordered triples (i, j, c), where *i* is the milestone the trajectory last crossed, *j* is the milestone the trajectory previously crossed, and *c* is the cluster (as defined by the centroid or histogram clustering methods) the trajectory sampled upon transition to *i*. As all the information needed to associate each sample trajectory with such states is readily available, the combination is straightforward and may be preformed at no significant additional cost. Furthermore, we note that such a combination may be advantageous as transition memory clustering has already been used to correct for the persistence of velocity memory[23] between transitions, while the previously discussed clustering methods may be applied to correct for the persistence of position memory.

2.3.4 Extracting Reaction Pathways

Perhaps the most significant feature of the MCM method is that it may be applied to further aid the milestoning method extract the reaction mechanism. The first step to doing so is defining the state space in a meaningful manner. If for example one is interested in the range of values a variable $u(x)$ assumes over the course of a reaction, a clustering procedure such as histogram clustering or centroid clustering should be used to define clusters corresponding to different intervals over u. To ensure the memory loss condition is satisfied, it may however still be necessary to define clusters with the aide of other variables as well. Within the context of a suitable state space, we may then determine the prevalence of different reactive paths. We define a reaction path as a sequence of states $\{\alpha_1,\alpha_2,...,\alpha_n\}$, where α_1 is contained in the set of states designated as reactants, α_n is in is contained in the set of states designated as products, and all intermediate states are neither in the products or the reactants. The absolute probability of any particular reaction path is then simply:

$$p_{\{\alpha_1,\alpha_2,...,\alpha_n\}} = T_{\alpha_n\alpha_{n-1}}...T_{\alpha_2\alpha_1}\int_0^\infty dtQ_{\alpha_1}(t), \tag{2.6}$$

where the integral gives the total volume of probability passing through α_1, and is computed with absorbing boundary conditions on all products states. It may be evaluated in closed form according to Eq. 3.8 in the next chapter.

The problem with a path by path analysis however is that there may be infinitely many such paths. We can still however answer general mechanistic questions by

considering categories of paths. For example, if we wish to know what conformational rearrangements must occur for the system to undergo a reaction, we may ask the question: What does the molecule generally look like when it *successfully* passes over a free energy barrier? In the context of MCM, we could answer this question by defining a milestone at the free energy barrier, and then finding for each MCM state μ on this milestone, the probability that μ is the last MCM state the system visits at the barrier top before going on to the products. Correspondingly, we may obtain the MTT, and the MFPT for paths associated with each such μ using the methods of 3.7.3, and an example in which we do so is given is in section 2.3.2.

2.3.5 Optimal number of clusters

Ideally, milestones would be sampled so well that the statistical noise seen in predictions of quantities like the MFPT is negligible, regardless of the number of clusters used on the milestones. In such a case, one could calculate the MFPT using a very large number of clusters—so large that using any more clusters would not change the MFPT's predicted value. When the number of sample transitions is bounded however, the number of sample transitions that describe any one state will become increasingly sparse as the number of clusters increase. To avoid statistical errors that may result from poor sampling, it is therefore reasonable to put a cap on the number of clusters. One way to establish this cap is to simply repeat the calculation of MFPT, using the same MD data set, over and over for an increasing number of clusters, ceasing calculations when the

observed statistical error becomes comparable to the predicted MFPT. Ideally, the predicted MFPT will converge to a value before the statistical error overwhelms. We note for each calculation of the MFPT, it's statistical error may be obtained using a Bayesian approach first pointed out by Vanden-Eijenden, et al.[22], and later implemented by Majek and Elber[4]. The only assumption required by this approach is that the lag time distributions are well described by exponential functions.

2.4 Application of MCM to a model process: Polymer reversal

We now illustrate MCM by applying it to the dynamics of a polymer confined to a cylindrical pore. Specifically, we are interested in computing the timescale for the polymer's end-to-end vector to reverse it's direction (Fig. 2.3).

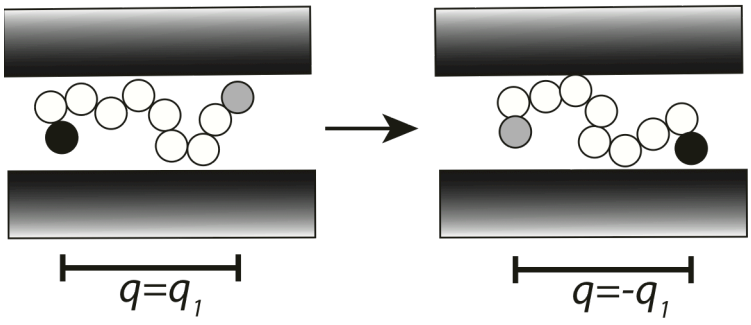


Figure 2.3: The reversal of a polymer, as monitored by the z coordinate of the polymer's end-to-end vector, $q(x) = z_N - z_1$.

This process has been used in a number of previous studies[36, 37] that sought to evaluate the effectiveness of kinetic models, and it poses a particularly good test to MCM

as it is a process in which position memory decays slowly with respect to the characteristic timescale of transitions between milestones.

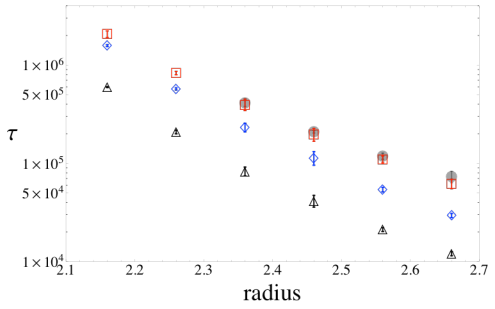
Noting that the polymer system possesses two equivalent free energy minima along our reaction coordinate, $q \equiv z_N - z_1$, at q^{mp} and $-q^{mp}$ (*mp* stands for most probable), we set out to calculate the mean first passage time and the mean transit time for the system to pass from one basin to the other. As with previous work[36], we consider a homo-polymer consisting of 35 monomers, leaving the details of our polymer and pore model to Appendix 2.5.2. Between the 2 free energy basins, we define 6 milestones which are spaced evenly along the reaction coordinate. Further details of how milestones are laid out are given in Appendix 2.5.3.

With our system thus defined, transitions between the milestones were sampled with the M1TM sampling procedure[23], and the statistics of transition were used to construct the transition probabilities and lag time distributions in an MCM state space in which both histogram clustering and transition memory clustering were used. The collective variable we use to perform Histogram Clustering is $u(x) = z_{3N/4} - z_{N/4}$. 40 MCM states were defined on milestone 1 (noting that it only has one connected milestone), and 80 MCM states were defined on each subsequent milestone. With these calculations, we obtained the dependence of the MFPT (Fig. 2.4a) and the MTT (Fig. 2.4b) on the pore radius. Consistent with previous theoretical[10, 43] and experimental findings[11, 13], we see that the MTT exhibits significantly weaker scaling with respect to the free energy barrier (controlled by the pore radius[36]) than does the MFPT, as the MTT

changes by less than a factor of 2 over our range of pore radii while the MFPT changes by almost 2 orders of magnitude.

While application of MCM significantly improves the milestoning calculations done here, it is important to emphasize that we have not necessarily performed conventional milestoning in an optimal fashion. For example, we did not seek to further optimize our choice of reaction coordinate. For a well chosen reaction coordinates, milestoning calculations can typically be brought at least to within a factor of two of the

a)



b)

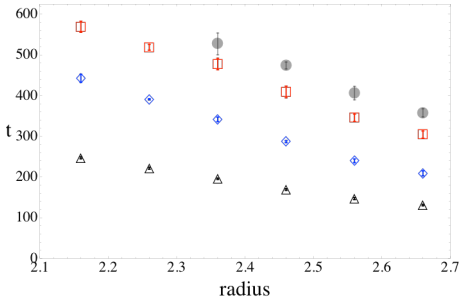


Figure 2.4: 2.4a and 2.4b gives the MFPT and MTT of polymer reversal as a function of the pore radius (in units of bond length) respectively. Circles are from MD calculations, red squares are from MCM calculations, blue diamonds are from M1TM calculations, and triangles are from conventional milestoning.

brute force MD predictions for simple systems where such comparison is possible. Examples of other reaction coordinate we could have used instead include a Steepest Descent Path(SDP), Minimum Energy Path, Minimum Free Energy Path, or a Finite Temperature String. As computing these reaction coordinates would have added a significant layer of complexity to the calculations done here, we chose instead to use a simple reaction coordinate that has been used in a previous polymer reversal study[36] and offers a clear measure of the reaction's progress. What is important to emphasize about the results shown in Fig. 2.4 is that they illustrate that when milestones are chosen such that the memory loss assumption is not satisfied, MCM can be used to significantly improve predictions of the kinetics.

That memory is not lost between the milestones of our system is clear from inspection of the transition probabilities. For example, when 8 MCM states are defined on each of milestones 2 and 3, we find that depending on the choices of MCM state β on milestone 2, the value of the transition probability $T_{\alpha\beta}$ (where α is fixed on milestone 3) can vary by as much as a factor of 38. If all memory was lost between transitions, $T_{\alpha\beta}$ would be invariant with respect to the choice of MCM β state on milestone 2. The persistence of memory between transitions was to be expected though since the autocorrelation time of $u(x)$ is comparable to the characteristic timescale of transitions between the milestones. Fig. 2.5 shows that the autocorrelation time of $u(x)$ — the characteristic timescale for the autocorrelation function

$$C_u(t) = \frac{\langle u(x(t))u(x(0))\rangle}{\langle u(x(0))^2\rangle} \tag{2.7}$$

to decay to a value of $1/e$ — is $\tau_u \simeq \frac{3}{5}\tau_q$, where τ_q is the autocorrelation time of the reaction coordinate. On the other hand, the characteristic timescale of transition between the intermediate milestones ranges from roughly $\tau_q/4$ when $r_{pore} = 2.76$ to roughly $\frac{3\tau_q}{2}$ when $r_{pore} = 2.16$.

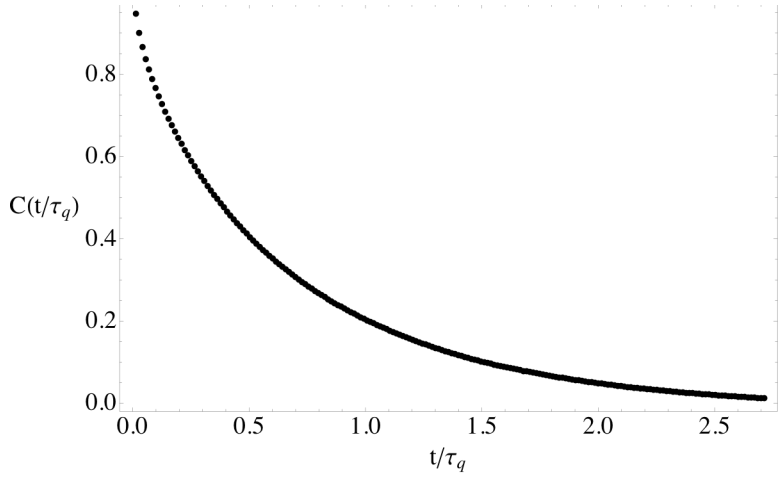


Figure 2.5: The autocorrelation function of $u(x) = z_{3N/4} - z_{N/4}$ for an *N*=35 bead polymer in a pore of radius 2.46 bond lengths. The time is normalized by autocorrelation time, τ_q, of the reaction coordinate, $q(x) = z_N - z_1$, illustrating $\tau_u \simeq \frac{3}{5}\tau_q$.

Before moving on, we make some technical remarks about our data in Fig. 2.4. First, calculations performed with half and twice the number of MCM states agreed within error of the results presented in Fig. 2.4, suggesting our predictions are converged

with respect to the number of MCM states. Second, we see that milestoning does achieve a substantial speed up compared to brute force, as seen is given in Fig. 2.6.

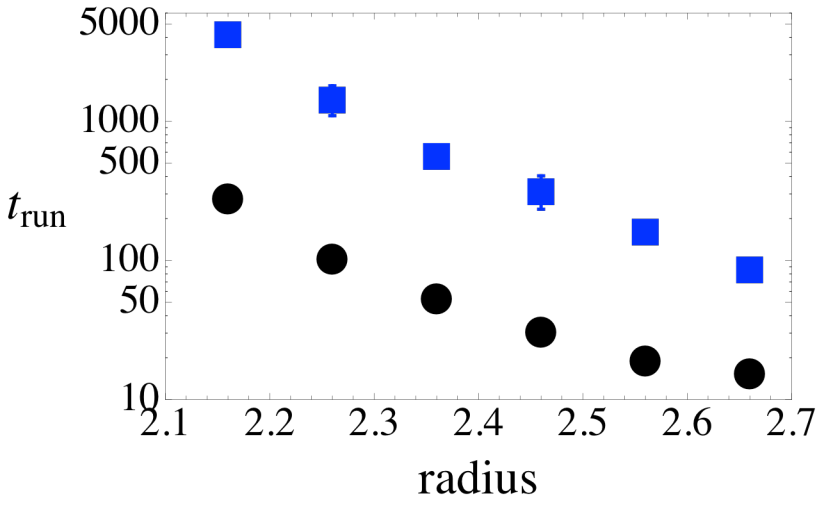


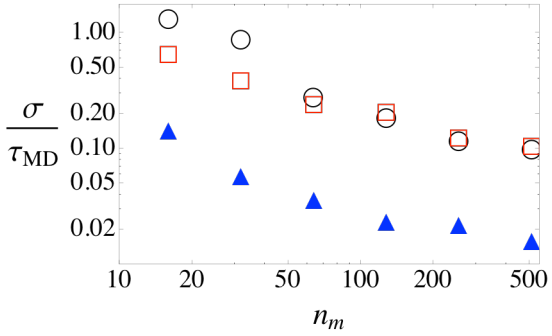
Figure 2.6: The runtime for calculations of the MFPT as a function of the pore radius. Circles are the runtimes of brute force MD calculations, and squares are the runtimes of the milestoning calculations. For the brute force MD calculations, we performed 4 independent trials, in which 162 reversal events were observed in each trial at each radius. 64 processors (16 Opteron quad-core 64 bit processors with core frequency of 2.3 GHz) were run in parallel with a master-slave load balancing algorithm. On the same hardware, the MCM calculations were obtained by requiring that all allowed transitions between the M1TM states were sampled by a minimum of 600 trajectories. Fig. 4 was generated from the same data of these calculations.

Further, we note that application of MCM on top of conventional milestoning increased the overall computation time by less than .01%.

2.4.1 Analysis of the statistical noise

In our example of polymer reversal, we were able to sample as many trajectories as were needed to keep the statistical noise of our MCM calculations small (relative to the predicted quantities) while still achieving a sizeable speed up compared to brute force molecular dynamics simulations. We now investigate how MCM performs as the number of sampled transitions is capped for various values of n_m, the minimum number of samples collected for any allowed *M1TM* transition. When n_m is large, we see in Fig. 2.7 that calculations using 32 MCM states per milestone yield better predictions of the MFPT than do conventional milestoning calculations and MCM calculations which use 8 MCM states per milestone.

a)



b)

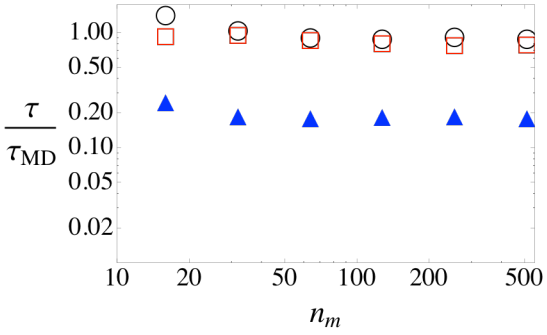


Figure 2.7: 2.7a gives the MFPT of reversal for a 35 bead polymer in a pore of radius 2.46 bond lengths, normalized by the true MFPT as determined by MD simulations (τ_{MD}), as a function of n_m. 2.7b gives the standard deviation of the MFPT (determined from 32 independent calculations), as a function of n_m. A total of 6 milestones were used in all calculations. Blue triangles correspond to conventional milestoning calculations. For circles, 16 MCM states are defined on milestone 1, 32 MCM states are defined on subsequent milestones. For red squares, 4 MCM states are defined on milestone 1, 8 MCM states are defined on subsequent milestones.

That is to say that the total error—considering both systematic (Fig. 2.7a) and statistical error (Fig. 2.7b)—is smallest when 32 MCM states per milestone are used,

considering just the three options presented. However, we see the trend that as the sample set becomes more diminished (i.e. as n_m decreases), the number of MCM states that minimizes the total error decreases. For example, for $n_m \leq 32$, the total error is smallest when 8 MCM states per milestone are used. Since the same MD data may be used to obtain the MFPT for both the case of 8 and 32 MCM states per milestone, it is fortunately not a problem to do the calculation for both choices and to then select the result with the ideal balance of statistical and systematic error, which, as discussed section 2.3.5, may be gauged by the ratio of the statistical error to the predicted MFPT.

2.4.2 Analysis of reaction paths

We would now like to learn about what structural rearrangements are necessary in order for polymer reversal to successfully occur. Namely, we ask, what values of $z_{\frac{3N}{4}} - z_{\frac{N}{4}}$ must the system assume at the free energy barrier before it may proceed to the products? To address this, we define 5 milestones in the system, where milestone 3 is placed at the free energy maxima located at $z_N - z_1 = 0$, and the others are spaced evenly over the transition region between the 2 free energy basins. We then compute the probability that each MCM state on milestone 3 is the *last* MCM state the system visited before moving on to the products (Fig. 2.8). We then compute the MTT (Fig. 2.9a) and the MFPT (2.9b) as a function of the last MCM state visited at the barrier top.

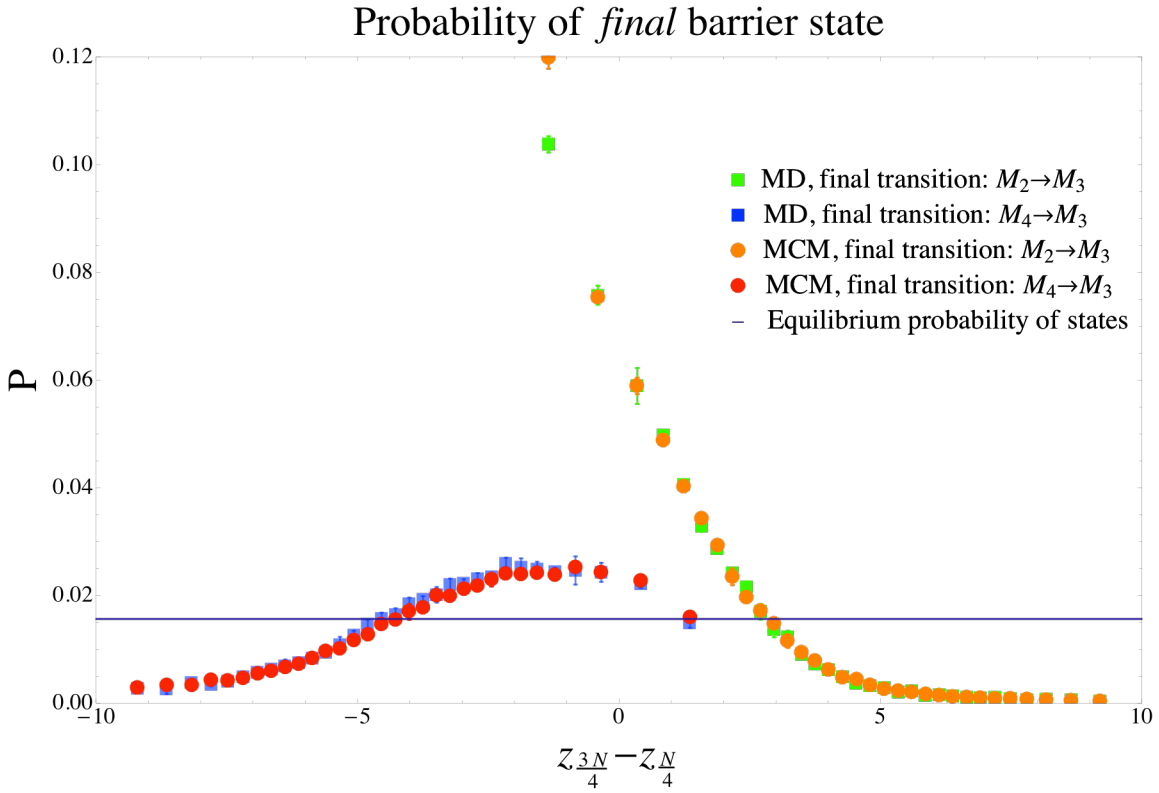
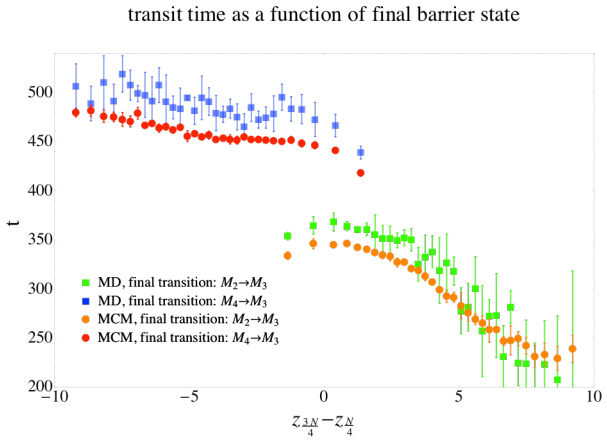


Figure 2.8: For each data point, the value of $z_{\frac{3N}{4}} - z_{\frac{N}{4}}$ gives the location of a cluster center for some MCM state (recall that each cluster is defined as an interval of $z_{\frac{3N}{4}} - z_{\frac{N}{4}}$ values), and the value of P is the probability of the MCM state being the last MCM state the system samples on milestone 3 before going on to the products. Green and blue squares were determined from long MD simulations, while the red circles and orange circles were determined from Eq. 3.38 in section 3.7.3. Note that since we have used both transition memory clustering and histogram clustering to define MCM states, there are 2 types of MCM states on the milestone 3: Those which can be accessed by trajectories making a transition from milestone 2 ($M_2 \to M_3$), those which can be accessed by trajectories making a transition from milestone 3 ($M_4 \to M_3$). The equilibrium probability of a trajectory at milestone 3 being in any state is constant (shown by the dark purple line). This is a consequence of using a fixed height histogram (fig. 2.2) to define clusters.

a)



b)

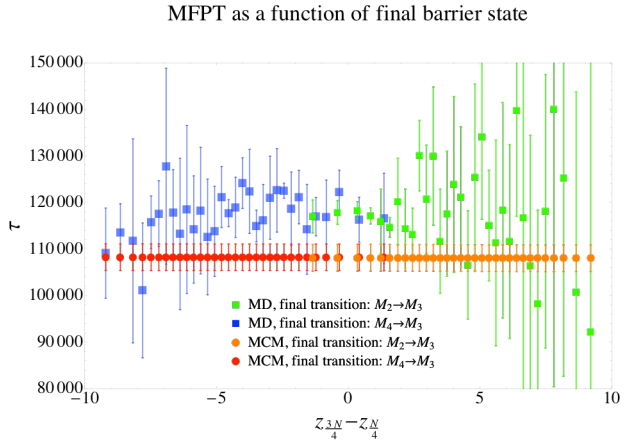


Figure 2.9: As with Fig. 2.8, for each data point, the value of $z_{\frac{3N}{4}} - z_{\frac{N}{4}}$ gives the location of a cluster center for some MCM state, and the value of t in 2.9a is the MTT exhibited by trajectories for which the final MCM state sampled at the barrier top is the one associated with the data point, while the value of τ in 2.9b is the MFPT for such trajectories. Eqs. 3.44 and 3.48 were used for MCM calculations of the MFPT and MTT respectively.

From Fig. 2.8, we see that before reversal events occurs, the internal coordinate, $z_{\frac{3N}{4}} - z_{\frac{N}{4}}$ typically exhibits a substantial contraction compared to the values seen in equilibrium, and that MCM captures this effect with surprising accuracy. From Fig. 2.9 we also see that MCM accurately captures the timescales associated with the various pathways through the barrier. Of potential experimental relevance is the fact that the MTT (Fig. 2.9a) exhibits a near one to one correspondence with the last barrier state visited over a wide range of values. This suggests that accurate measurements of the transit time could in principle be directly associated with the reaction mechanism. The MFPT on the other hand is roughly constant across all reaction paths, indicating measurements of it would give very little mechanistic information. The relative constancy of the MFPT does not come as a major surprise because the time our system spends on any successful reaction path is orders of magnitude smaller than the time spent fluctuating about the reactants.

2.5. Concluding remarks

We have introduced a simple generalization of milestoning, MCM, which may be applied on top of any the current milestoning methods without changing how transitions are sampled between milestones and without changing any of the fundamental milestoning equations. The only difference the method introduces is a post processing

stage in which trajectories exhibiting transitions are associated with MCM states based on their coordinates at the times of transition, and or the identity of the milestone they previously visited. We have suggested a number of clustering procedures for defining MCM states, and have shown in particular that when Transition Memory and Histogram Clustering are used in combination with each other, MCM accurately recovers the dynamics of a system ill suited for conventional milestoning calculations.

At this point it is worth noting that the method of Directional milestoning[4, 6-9, 44] (DiM) provides a means of ensuring milestones are spaced far enough apart such that memory loss may occur between transitions. However, there are still several advantages to performing DiM with MCM. First, as discussed in section 2.2, MCM may help with the extraction of the reaction mechanism. Second, in DiM calculations, it is necessary to specify the states of the system (i.e. directional milestone) before any calculations are done. If the memory loss condition is not sufficiently satisfied for the choice of milestones, the calculation, including MD simulations, must be repeated. MCM on the other hand does not require specification of the state space until after all transitions have been sampled. Even then, one is not limited to specifying a single state space. If the initial choice of state space proves inadequate (e.g. calculations of the MFPT show significant disagreement with the MFPT obtained when a finer version of the state space is used), one may simply try calculations in a different state space, without having to repeat any molecular dynamics simulations.

A number of important questions regarding this method still remain. What problems are not well suited for MCM? What clustering techniques offer the best balance of computational efficiency and accuracy for what classes of MD problems? Answers to these questions will only emerge after MCM has been applied to a greater variety of systems.

2.6 Appendix

2.6.1 Validation of the memory loss assumption

Here, we show that Eqs. 2.1 and 2.2 may be obtained from Eq. 2.4 in the general MCM state space. We closely follow the derivation of Kreuzer, Elber, and Moon[6], which established Eqs. 2.1 and 2.2 in context of conventional milestoning.

We start by defining $Q_\alpha\left(x_\alpha,t\right)$ as the probability density that the system assumes configuration x_α at time *t*. In terms of initial probabilities and the transition kernel defined in section 2.2, $K_{\alpha\beta}\left(x_\alpha,x_\beta;t'\right)$, it may be written as

$$Q_\beta\left(x_\beta,t\right)=p_\beta\left(x_\beta,0\right)\delta\left(t^+\right)+\sum_{\alpha\in\bar{\beta}}\int_{S_\alpha}dx_\alpha\int_0^t dt'Q_\alpha\left(x_\alpha,t'\right)K_{\alpha\beta}\left(x_\alpha,x_\beta;t'\right), \tag{2.8}$$

where $p\left(x_\beta,0\right)$ is the initial probability that the system assumes configuration x_β, and S_α is the region of phase space occupied by the cluster that defines state α.

Before proceeding, we observe some definitions:

$$Q_\beta\left(t\right)\equiv\int_{S_\beta}dx_\beta Q\left(x_\beta,t\right), \tag{2.9}$$

$$p_\beta(t) \equiv \int_{S_\beta} dx_\beta p(x_\beta, t) \ , \tag{2.10}$$

$$K_{\alpha\beta}(t \mid x_\alpha) \equiv \int_{S_\beta} dx_\beta K(x_\alpha, x_\beta, t), \tag{2.11}$$

where $K_{\alpha\beta}(t \mid x_\alpha)$ is the conditional probability density that a trajectory initialized at x_α in α makes a transition directly to β after a time *t*. When $K_{\alpha\beta}(t \mid x_\alpha)$ has no dependence on x_α (e.g. when the memory loss assumption is valid), we just write it is $K_{\alpha\beta}(t)$.

Now, integrating both sides of (A1) over the cluster that defines state β, and applying the memory loss assumption Eq. 2.4, we obtain

$$\int_{S_\beta} dx_\beta Q_\beta(x_\beta, t) = \int_{S_\beta} dx_\beta p_\beta(x_\beta, 0)\delta(t^+) + \sum_{\alpha \in \bar{\beta}} \int_{S_\beta} dx_\beta \int_{S_\alpha} dx_\alpha \int_0^t dt' Q_\alpha(x_\alpha, t') K_{\alpha\beta}(x_\beta; t - t') . \tag{2.12}$$

Using Eqs. 2.9-2.11, Eq. 2.12 simplifies to Eq. 2.2:

$$Q_\beta(t) = p_\beta(0)\delta(t^+) + \sum_{\alpha \in \bar{\beta}} \int_0^t Q_\alpha(t') K_{\alpha\beta}(t - t') dt' .$$

Eq. 2.1 follows immediately from the validity of Eq. 2.2.

2.6.2 Polymer and pore models

Our polymer model consists of *N* =35 beads of mass m connected by *N*-1 springs. The pairwise interaction potential between adjacent beads is:

$$V_{bond}(r) = k_{bond}(r - \sigma)^2 / 2 \tag{2.13}$$

The non-bonded beads interact via a repulsive Leonard-Jones potential representing excluded volume effects:

$$V_{non-bonded}(r) = 4\varepsilon\left[(\sigma / r)^{12} - (\sigma / r)^{6}\right]\theta(2^{1/6}\sigma - r). \qquad (2.14)$$

Here σ is the equilibrium bond distance, r is the distance between bead centers, ε is the characteristic energy scale, $k_{bond} = 100\varepsilon / \sigma^2$, and θ is the Heaviside step function that truncates the attractive portion of the Lennard Jones potential.

Further, each bead is subject to a repulsive interaction with the pore:

$$V_{pore}(r) = 1000\varepsilon\left(\frac{\left(x_i^2 - y_i^2\right)^{1/2} - r_{pore}}{\sigma}\right)^4 \theta\left(\left(x_i^2 - y_i^2\right)^{1/2} - r_{pore}\right), \qquad (2.15)$$

where r_{pore} characterizes the radius of the pore, and θ denotes a Heavyside step function. An equivalent pore in which the boundary is an impenetrable hard-wall has a radius: $R = r_{pore} + 0.66$, as determined in previous literature[36]. All reported radii are reported in terms of R.

The dynamics of each bead is governed by the Langevin equation:

$$m\ddot{r}_i(t) = -\frac{\partial V}{\partial r_i} - \xi\dot{r}_i(t) + f(t) \qquad (2.16)$$

where $r_i(t)$ is the position of the bead, V is the total interaction potential, ξ is the friction coefficient, and $f(t)$ is a random force satisfying the fluctuation dissipation theorem. Our data is reported in reduced units, where the units of length and energy are σ and ε.

In our simulations, the unit of time is $t = \sqrt{m\sigma^2 / \varepsilon}$, the temperature $T = 1.0\varepsilon / k_B$, and our friction coefficient is $\xi = 2.0(\sigma^2 / m\varepsilon)^{-1/2}$.

2.6.3 Milestones in polymer reversal

To be consistent with previous work[36], our first milestone is positioned at q_1, the reaction coordinate value that satisfies the condition: $P(q_1) = \frac{P(q^{mp}) + P(0)}{2}$, where $P(q)$ is the equilibrium probability of observing the system in q. Then the remaining milestones are positioned at locations: $q_i = q_1 - \frac{2q_1}{M-1}(i-1)$, where *M* is the total number of milestones used in the calculation.

Chapter 3. Computation of transit times using the milestoning method with applications to polymer translocation

3.1 Abstract

Milestoning is an efficient approximation for computing long-time kinetics and thermodynamics of large molecular systems, which are inaccessible to brute-force molecular dynamics simulations. A common use of milestoning is to compute the mean first passage time (MPFT) for a conformational transition of interest. However, the MFPT is not always the experimentally observed timescale. In particular, the duration of the transition path, or the mean transit time (MTT), can be measured in single-molecule experiments, such as studies of polymers translocating through pores. Here we show how to use milestoning to compute transit times and illustrate our approach by applying it to the translocation of a polymer through a narrow pore. In doing so, apart from showing our method to be highly accurate, we also confirm a theoretical prediction stating that the transit time obeys a power law scaling relationship $t \propto N^{\alpha}$ with respect to the polymer chain length, N, where the scaling exponent, α, is the same as the scaling exponent that governs the chain length dependence of another timescale observed in translocation studies, the escape time. In addition to giving a milestoning procedure for obtaining the transit time, we also show how milestoning calculations in general may be made more efficient through exploitation of time reversal symmetry.

3.2 Introduction

Many biomolecular processes, such as protein or RNA folding and large-scale conformational rearrangements in proteins, occur on time scales prohibitive of brute force molecular dynamics (MD) simulations. A number of approximate methods have been proposed to overcome this hurdle[4, 5, 20, 22, 23, 25, 26, 28-33, 35, 45]; of those, milestoning is a robust and highly efficient approach for extracting both the kinetics and thermodynamics of large biomolecular systems, often providing computational gains that span many orders of magnitude [9, 24]. Milestoning is especially useful for systems undergoing diffusive dynamics such that the exponential kinetics assumption, adopted by other related methods such as Transition Path Sampling[46], Transition Interface Sampling[30], and Partial Path Transition Interface Sampling[26], is unnecessary.

Milestoning provides a statistical procedure, whereby the sought after long-time trajectories of a molecular system are recreated by gluing together shorter trajectory fragments that originate and terminate on milestones (predefined regions of configuration space, typically hypersurfaces) placed between the reactants and products of a reaction. The method was initially developed for the purpose of obtaining timescales, such as the mean first passage time (MFPT), of processes governed by a single reaction coordinate[5, 20]. With recent developments however, it has been extended to complex systems possessing multiple reaction pathways[4, 22], providing both reaction timescales and mechanistic information such as maximum-flux pathways[47].

Most of the milestoning studies, as well as other related methods, have focused on

computing the first passage time from some well defined state to another. This choice is closely connected to the experimental observables: Indeed, the first passage time for the transition from one basin of attraction (say A) to another (B) equals the inverse of rate coefficient for the $A \to B$ reaction, a quantity that is directly measurable by many experimental techniques. Recent single-molecule studies, however, have brought to spotlight a different timescale: the duration of a transition event, commonly referred to as a transit time. Notable examples include DNA and polypeptide translocation, as measured by nanopore force spectroscopy [48-59], the duration of protein / RNA folding events as measured by single-molecule Forester Resonance Energy Transfer (sm-FRET) studies[10-12], and the duration DNA / RNA folding events by single-molecule force spectroscopy[13, 14].

The distinction between first passage times and transit times is illustrated in Fig. 3.1 Trajectories that represent samples of these times both start in a "reactant state", A, and later terminate in a, "product" state, B. However, trajectories sampling the first passage time will exhibit multiple failed attempts, where they enter the transition region lying between A and B and exit back to A, having failed to accomplish the transition. In contrast, trajectories which sample the transit time illustrate successful attempts at the reaction, as they enter the transition region and exit into B without ever going back to A.

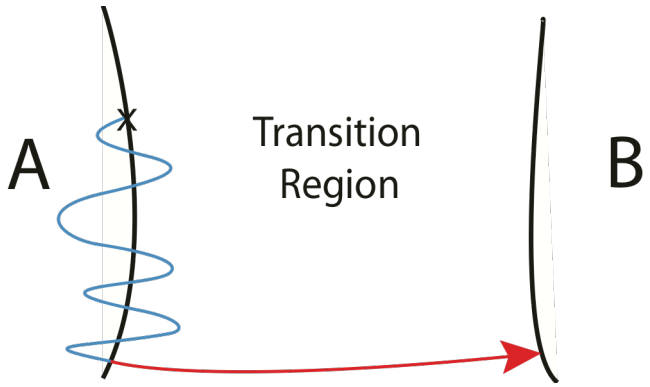


Figure 3.1: A schematic diagram of a trajectory that passes from the reactants region of conformation space to the products. The blue line represents the trajectory while it is fluctuating between the reactants and the transition region. The red line represents the reactive portion of trajectory and it's time duration gives a sample of the transit time whereas the complete trajectory would be a sample of the first passage time.

Given the increasing interest in the finer details of transition paths encountered in biomolecular rearrangements and, particularly, in the duration of such paths, a number of theoretical studies of transit times have emerged recently[43, 60-64]. The scope of the methods discussed in these studies is limited however to systems whose reaction dynamics in the transition region may effectively be described by a few simple properties of the energy landscape there, such as the frequency of the unstable mode at a barrier top, and the height of the energy barrier. For large biomolecular systems possessing multiple complex reaction pathways, such a reduced description may fail to capture the dynamics of the system. Molecular dynamics simulations that account for the relevant details of the energy landscape therefore represent the most general approach to the transit time.

As a brute force MD simulation can be exceedingly costly, we here give a milestoning method for computing transit times.

As achieving computational efficiency in molecular simulations is our general purpose here, we also take the time to discuss how milestoning calculations in general may be expedited. Using an argument based on time reversal symmetry, we show how to extract better statistics from the molecular dynamics simulations performed in milestoning.

To test this method, we use it to compute the time it takes a long polymer chain to traverse a narrow pore[65-71] and find that it agrees well with brute-force simulations and accurately reproduces the scaling[65-72] of this time with chain length, despite significant memory effects[73-75] present in this problem. The rest of this paper is organized as follows. In section 3.3 we discuss the milestoning method and show how it can be used to obtain transit times. In section 3.4, we briefly discuss how to improve the efficiency of milestoning in general via considerations of time reversal symmetry. In section 3.5 we apply this milestoning algorithm to polymer translocation, and Section 3.6 concludes with closing remarks.

3.3 Transit times from the milestoning method

3.3.1 Milestoning

We start by reviewing the standard milestoning method and the assumptions it makes. In the method, one first defines milestones – regions of conformation space (typically hypersurfaces) that the system is likely pass through over the course of its time evolution. The sequence of milestones that a long trajectory visits, $\{\alpha_1,\alpha_2,...,\alpha_n,...\}$, is then approximated by a discrete Markov chain described by the discrete-time master equation of the form

$$p_\alpha^{(n)} = \sum_\beta T_{\alpha\beta} p_\beta^{(n-1)}, \tag{3.1}$$

where $p_\alpha^{(n)}$ is the probability of making a transition to milestone α at step *n*, and $T_{\alpha\beta}$ is the conditional probability to make a transition to α provided that the system has currently at milestone β. One further assumes the corresponding sequence of lag time between transitions to milestone α_n and α_{n+1} is determined by a probability distribution, $\pi_{\alpha_{n+1},\alpha_n}(t)$ which only depends on these 2 milestones. The conditions under which these assumptions are rigorously satisfied have been analyzed in detail previously[21]. In practice it has been found that milestoning gives a good approximation of the system's dynamics when velocity memory is lost between successive milestones[5, 20]; and even

then, a version of milestoning, M1TM[23], has given accurate predictions of the kinetics of a system in which velocities of trajectories were highly correlated between successive milestones.

In general, the evolution of the system in the reduced milestoning state space may be recovered through a coupled set of integro-differential equations[5, 9, 20]

$$p_\alpha(t) = \int_0^t Q_\alpha(t') \left[1 - \int_0^{t-t'} \sum_\beta K_{\beta\alpha}(t'') dt'' \right] dt' \tag{3.2}$$

$$Q_\alpha(t) = p_\alpha(0)\delta\left(0^+\right) + \sum_\beta \int_0^t Q_\beta(t') K_{\beta\alpha}(t - t'') dt' \tag{3.3}$$

$$K_{\alpha\beta}(t) \equiv T_{\beta\alpha}\pi_{\beta\alpha}(t). \tag{3.4}$$

Here, $p_\alpha(t)$ is the probability that the system is in state α at time *t*, and $Q_\alpha(t)$ is the flux of probability into α at time *t*. We note that the sub-indices are transposed from the left hand side to the right hand side of Eq. 3.4 to be consistent with conventions for these symbols in the literature[9, 23].

However, integrating these equations would be an overkill if one is only interested in the mean first passage time of a process, as is often the case, since more efficient procedures for obtaining it have been given[4, 5, 21, 23]. In the next two subsections, we give procedures for obtaining the probability distribution of transit times and the mean transit time (MTT).

3.3.2 Probability distribution of transit times

Below we derive the probability distribution of transit times and the MTT assuming conventional milestoning in the special case where a reaction coordinate is used to order the milestones. For more general milestoning schemes, the probability distribution of transit times and the MTT are given in Appendices 3.7.1 and 3.7.2 respectively. Appendxix 3.7.3 further gives a procedure for obtaining the MTTs associated with different reaction pathways in the system.

Consider the ensemble of transition paths that cross milestone 1 at some time and later arrive at milestone M without ever returning to milestone 1. Each of those transition paths can be split into two segments. The first segment corresponds to a transition from milestone 1 to milestone 2 and the second segment describes the rest of the transition path (Fig. 3.2).

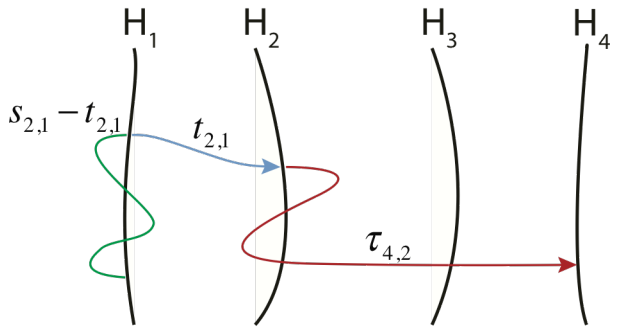


Figure 3.2: A schematic diagram of a trajectory that passes from milestone 1 to milestone 4. The duration of the reactive portion of this trajectory, $t_{2,1}+\tau_{4,2}$ represents a sample of the transit time for this event. The duration of the blue portion of the path represents a sample of the transit time for the event that the system goes from milestone 1 to 2. The duration of the green and blue portions of the path combined represents a sample of the lag time, $s_{2,1}$.

Accordingly, the transit time of a particular trajectory, $t_{M,1}$ may be written as $t_{M,1}=t_{M,2}+t_{2,1}$, where $t_{2,1}$ is the time for a trajectory in an ensemble to reach milestone 2 from 1, conditional upon never returning to milestone 1, and $t_{M,2}$ is the time for the trajectory to then go on to milestone *M* from milestone 2, also conditional upon not returning to milestone 1. If $t_{2,1}$ and $t_{M,2}$ are statistically independent, the probability distribution of transit times will be

$$P_{M,1}(t)=\int_0^\infty dt' u_{M,2}(t-t')P_{2,1}(t'), \tag{3.5}$$

where $P_{2,1}(t)$ and $u_{M,2}(t)$ are the probability distributions of $t_{2,1}$ and $t_{M,2}$ respectively. The first of these two distributions is evaluated directly from the sample set of trajectories observed to go from milestone 1 to 2, as each trajectory in this set yields a sample of $t_{2,1}$.

The probability distribution of times $t_{M,2}$, $u_{M,2}(t)$, can be found from Eq. 3.3 as

$$u_{M,2}(t) = \frac{Q_M(t)}{p_M(\infty)}, \tag{3.6}$$

where $Q_M(t)$ is calculated with the condition that the system starts on milestone 2 and absorbing boundary conditions[62] are imposed both at milestones 1 and M. The absorbing boundary condition at the first milestone ensures that trajectories that return to milestone 1 do not contribute to $Q_M(t)$. Eq. 3.6 is therefore normalized[61, 62] by $p_M(\infty)$, the total probability for the system to be absorbed by milestone M –i.e. the splitting probability.[76]

Computing the denominator of Eq. 3.6 is a simple matter. Denoting the Laplace transform of $Q_M(t)$ as $\tilde{Q}(\alpha)$, we observe that the probability for the system to be absorbed by milestone M is

$$p_M(\infty) = \int_0^\infty dt' Q_M(t') = \tilde{Q}_M(0), \tag{3.7}$$

where the first identity holds because milestone M is absorbing.

We now use the identity[5]

$$\tilde{Q}_M(\alpha) = \mathbf{p}_f^T \left[\mathbf{I} - \tilde{\mathbf{K}}_d(\alpha)\right]^{-1} \mathbf{p}_i, \tag{3.8}$$

where $\mathbf{p}_i$ is the vector corresponding to the initial milestone (in this case $\mathbf{p}_i = \mathbf{e}_2 = \langle 0,1,0.....,0 \rangle^T$), $\mathbf{p}_f^T$ is the unit vector corresponding to milestone M (i.e.

$\mathbf{p}_f^T = \mathbf{e}_M = \langle 0,0,.....0,1\rangle^T$), $\mathbf{I}$ is the M by M identity matrix, and $\tilde{\mathbf{K}}_d(\alpha)$ is the Laplace transform of the matrix $\mathbf{K}_d(t)$, which has elements $\left(\mathbf{K}_d(t)\right)_{\alpha\beta} = \pi_{\alpha\beta}(t)\hat{T}_{\alpha\beta}$. Here, $\hat{\mathbf{T}}$ is the modified transition probability matrix specified by

$$\hat{T}_{\alpha\beta} = \begin{cases} 0 & \text{if } \beta = 1 \text{ or } \beta = M \\ T_{\alpha\beta} & \text{otherwise} \end{cases} \tag{3.9}$$

defined in accordance with the absorbing boundary conditions. Observing that $\tilde{\mathbf{K}}_d(0) = \hat{\mathbf{T}}$, Eqs. 3.7 and 3.8 give

$$p_M(\infty) = \mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2 . \tag{3.10}$$

Using Eq. 3.7 in Eq. 3.6 thus gives

$$u_{M,2}(t) = \frac{Q_M(t)}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2} . \tag{3.11}$$

Using Eq. 3.11 in Eq. 3.5, we finally arrive at

$$P_{M,1}(t) = \frac{\int_0^\infty dt' P_{2,1}(t') Q_M(t-t')}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2} . \tag{3.12}$$

3.3.3 The Mean transit time

We now give a simple, computationally tractable expression for the MTT, $\langle t_{M,1} \rangle$. We start by expressing it as the sum:

$$\langle t_{M,1} \rangle = \langle \tau_{M,2} + t_{2,1} \rangle = \langle \tau_{M,2} \rangle + \langle t_{2,1} \rangle , \tag{3.13}$$

where the time, $\langle t_{2,1} \rangle \equiv \int_0^\infty t P_{2,1}(t) dt$, can be computed explicitly from observed transitions from milestone 1 to 2. The time, $\langle \tau_{M,2} \rangle \equiv \int_0^\infty t u_{M,2}(t) dt$ can be equivalently expressed as $\langle \tau_{M,2} \rangle = -\left. \frac{d\tilde{u}_{M,2}}{dz} \right|_{z=0}$, where $\tilde{u}_{M,2}(z)$ is the Laplace transform of $u_{M,2}(t)$, given in Eq. 3.11. Evaluating $\tilde{u}_{M,2}(z)$, we get

$$\tilde{u}_{M,2}(z) = \mathcal{L}\left[u_{M,2}(t)\right] = \mathcal{L}\left[\frac{Q_M(t)}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2}\right] = \frac{\tilde{Q}_M(z)}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2}. \tag{3.14}$$

Using Eq. 3.8 to replace $\tilde{Q}_M(z)$, we have

$$\tilde{u}_{M,2}(z) = \frac{\mathbf{e}_M^T \left[\mathbf{I} - \tilde{\mathbf{K}}_d(z)\right]^{-1} \mathbf{e}_2}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2}. \tag{3.15}$$

Therefore $\langle \tau_{M,2} \rangle$ is given by:

$$\langle \tau_{M,2} \rangle = -\left. \frac{d\tilde{u}_{M,2}}{dz} \right|_{z=0} = \left. \frac{\mathbf{e}_M^T \left[\mathbf{I} - \tilde{\mathbf{K}}_d(z)\right]^{-1} \left[\frac{d}{dz} \tilde{\mathbf{K}}_d(z)\right] \left[\mathbf{I} - \tilde{\mathbf{K}}_d(z)\right]^{-1} \mathbf{e}_2}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2} \right|_{z=0}, \tag{3.16}$$

where we have used $\frac{d}{dz}\mathbf{A}^{-1} = \mathbf{A}^{-1} \frac{d\mathbf{A}}{dz} \mathbf{A}^{-1}$ to take the derivative of $\left[\mathbf{I} - \tilde{\mathbf{K}}_d(z)\right]^{-1}$.

Using

$$\left. \frac{d}{dz} \left(\tilde{\mathbf{K}}_d\right)_{\alpha\beta} \right|_{z=0} = \left. \frac{d}{dz} \left(\int_0^\infty e^{-\alpha t} \pi_{\alpha\beta}(t) \hat{\mathbf{T}}_{\alpha\beta} dt \right) \right|_{z=0} = \hat{\mathbf{T}}_{\alpha\beta} \int_0^\infty t \pi_{\alpha\beta}(t) dt = \hat{\mathbf{T}}_{\alpha\beta} \langle \tau_{\alpha\beta} \rangle \tag{3.17}$$

and that $\tilde{\mathbf{K}}_d(0) = \hat{\mathbf{T}}$ in Eq. 3.14, we get

$$\left\langle \tau_{M,2} \right\rangle = \frac{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \hat{\mathbf{L}} \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2}{\mathbf{e}_M^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_2}, \tag{3.18}$$

where $\hat{\mathbf{L}}$ is the matrix with elements $\hat{L}_{\alpha\beta} \equiv \left\langle s_{\alpha\beta} \right\rangle \hat{T}_{\alpha\beta}$, with $\left\langle s_{\alpha\beta} \right\rangle$ defined as the mean lag time: $\left\langle s_{\alpha\beta} \right\rangle \equiv \int_0^\infty s\pi_{\alpha\beta}(s)ds$. All other quantities in Eq. 3.18 were defined above.

Defining $\mathbf{A} \equiv \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1}$, and using Eq. 3.18 in Eq. 3.13, we obtain the MTT as

$$\left\langle t_{M,1} \right\rangle = \frac{\mathbf{e}_M^T \mathbf{A} \hat{\mathbf{L}} \mathbf{A} \mathbf{e}_2}{\mathbf{e}_M^T \mathbf{A} \mathbf{e}_2} + \left\langle t_{2,1} \right\rangle. \tag{3.19}$$

While our expression Eq. 3.12 for the probability distribution of transit times required an additional assumption that the times $t_{2,1}$ and $t_{M,2}$ are statistically independent, Eq. 3.19 is valid even when these quantities are statistically dependent, as we can see from Eq. 3.13. Finally we note that all higher moments of the transit time distributions are given by the identity $\left\langle t_{M,1}^n \right\rangle = (-1)^n \left. \frac{d^n \tilde{P}_{M,1}}{dz^n} \right|_{z=0}$, where $\tilde{P}_{M,1}(z)$ is the Laplace transform of $P_{M,1}(t)$. To obtain closed form expressions for these moments however, it is necessary to assume $t_{\beta\alpha}$ and τ_β are statistically independent.

3.4 Using time reversal symmetry to improve sampling efficiency

The underlying assumptions of milestoning are satisfied if the physical process in question loses memory in the time between transitions. For memory loss to occur, physically, the first hitting point distribution (the probability distribution over the points

in phase space a trajectory may sample upon transition to a new state) a trajectory samples on any given state must be independent of the configuration it sampled at the time of it's previous transition. Clearly then, the first hitting point distribution on state β is the probability distribution sampled by trajectories in the equilibrium ensemble of the system upon arrival to β. If it were not, the memory loss assumption would be violated when the system is in equilibrium. The first hitting point distribution on β may therefore be sampled by selecting trajectories from the equilibrium ensemble of the system that are seen to go from any other state α, to β. The statistics of transition for β may be computed by following each such trajectory until it arrives at another state, say γ –i.e. $T_{\gamma\beta}$ and $\pi_{\gamma\beta}(t)$ may be computed from observation of these transitions. But, since these trajectories are members of the equilibrium ensemble of the system, their time reversed paths are also members of the equilibrium ensemble with a statistical weight equivalent to their forward in time counterparts. Therefore, each trajectory seen go from α, to β, to γ can be reversed to give a trajectory that goes from γ to β, to α, thereby also providing a sample transition from β to α (Fig. 3.3).

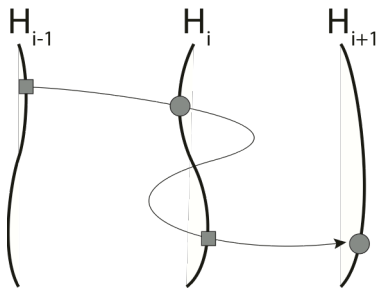


Figure 3.3: A schematic representation of a trajectory that starts on milestone *i*-1 and passes to *i*+1, assuming a conventional milestoning state space. The time elapsed between first hitting milestone *i* and terminating on *i*+1 yields a sample of $s_{i+1,i}$, the lag time for the system to transition from *i* to *i*+1. The portion of the path corresponding to this time lies within the circles. The time reversed trajectory however yields a sample of $s_{i-1,i}$, which corresponds to the portion of the trajectory that lies between the squares.

3.5 Polymer translocation

To illustrate our method, we calculate the time it takes a polymer to diffuse though a narrow pore. This quantity, also known as the polymer translocation time, is an example of a transit time because experimental studies of polymer translocation directly measure the time the polymer dwells inside the pore. [48-59] Computation of polymer translocation time presents a suitable test for our method for two reasons: First, threading of a polymer through a pore is a highly non-Markovian process that exhibits significant memory[73-75, 64] thereby challenging the milestoning memory loss approximation. Second, much of the recent work has focused on computing the scaling laws of various translocation timescales as a function of polymer length[65-75, 77]. However, in MD studies that sought to report the scaling behavior of the transit time, the "escape time" —which is

the time for a polymer that has translocated halfway (i.e. half of it's monomers are on one cis side of the pore, and the other half are on the trans side) to completely exit the pore to either side—was used as a proxy, based on theoretical claims[4365, 69] that the mean escape time should exhibit the same scaling behavior as the MTT. We therefore seek to directly compute the scaling relationship between the polymer chain length and the transit time, and then test the claim that the transit and escape times have the same scaling behavior.

Similarly to previous studies[67, 68, 70, 71], we consider the translocation of a bead-and-spring model homopolymer through a narrow constriction (see Appendix 3.7.4 for the details of the model). The process begins with an end monomer, which we call monomer 1, on the *trans* side of the pore and all other monomers starting on the *cis* side of the pore. The process ends when all monomers reach the trans side of the pore (Fig. 3.4).

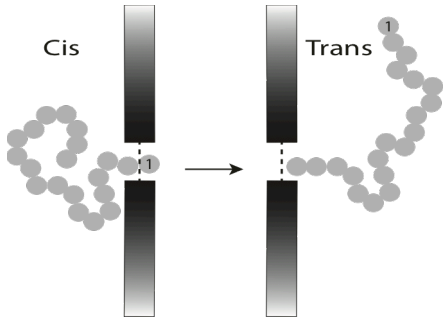


Figure 3.4: A homopolymer translocating through a narrow pore. The dashed line represents a reflecting boundary that applies only to monomer 1.

In simulations where no biasing force is present to drive the polymer to the *trans* side of the pore (which we do not include) it is conventional[66, 77] to impose a reflecting boundary condition on monomer 1 that confines it to the trans side of the pore. The average total time for the process to complete is the MFPT, and the average time for a trajectory to complete translocation after last hitting the reflecting boundary is the transit time. We note the MFPT of translocation is a somewhat artificial quantity because a reflecting boundary applied to just monomer 1 is not physical. Nevertheless, this time has been previously used by others [66, 77] as a measure of the translocation time and so we compute it for the purpose of comparison.

The progress of the translocation process is quantified by the reaction coordinate , $q(x)$, equal to the number of monomers on the *trans* side of the pore. We then define milestones 1 through *M* as the regions in conformation space H_1 through H_M, where:

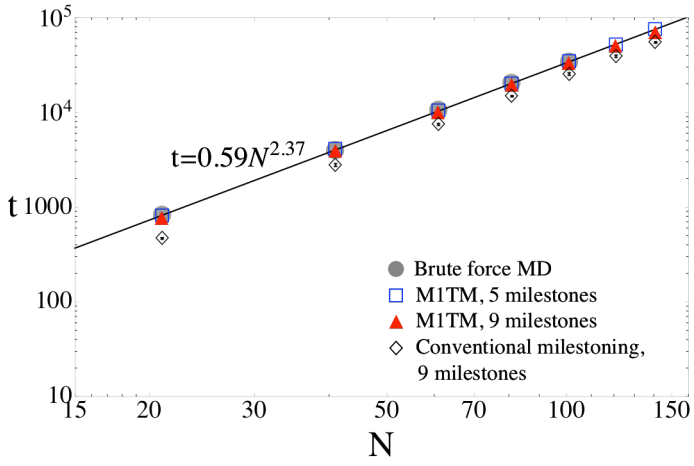
$$H_i \equiv \{x : q(x) = q_i\}. \tag{3.20}$$

Here q_i is an integer value on the interval [0, *N*], and *x* is a point in configuration space. We take $q_1 = 0$ and $q_M = N$ so that these milestones correspond to the initial and final states of the translocation process. The remaining milestones are spaced evenly along the reaction coordinate between q_1 and q_M.

The results of our calculations for the MTTs and MFPTs are reported, respectively, in Figs. 3.5a and 3.5b. Note that, while, for simplicity, Section 2 presented our method under the assumptions of standard milestoning, the actual calculations

employed both conventional milestoning and the M1TM method[23] described in Appendices A and B. The observations of section 3.4 regarding time reversal symmetry were also employed to boost the efficiency of our calculations.

a)



b)

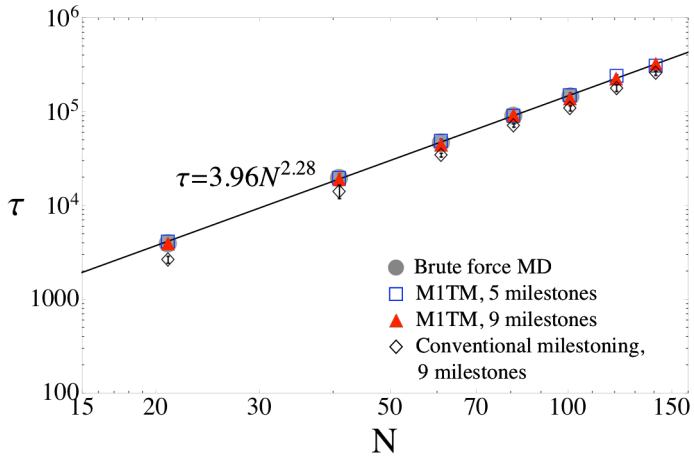


Figure 3.5: Mean transit time (a) and MFPT (b) plotted as a function of chain length N. The straight lines are least squares fits of the 5 milestone M1TM data to functions of the form $f(N) = bN^{\alpha}$.

As seen in Fig. 3.5, M1TM calculations were more accurate for both the MTT and the MFPT. In particular, calculations with 5 and 9 milestones were within 7% of the brute force predictions, while the 17 milestone calculations (not shown in Fig. 3.5) did not deviate by more than 20% . In contrast, conventional milestoning calculations using 9 milestones were between 60% and 72% of the brute force transit times. This finding is

not surprising given that polymer translocation is notoriously non-Markovian.[73-75] , suggesting the conventional milestoning assumption of memory loss assumption is not well satisfied in our simulations. M1TM, which partially accounts for the memory effects, performs considerably better.

Power-law fits of our mean transit and mean first passage times, as a function of chain length *N*, yield $\tau \propto N^{\alpha}$, with a scaling exponent of $\alpha = 2.28 \pm .02$ for the mean first passage time and $\alpha = 2.37 \pm .01$ for the MTT, as determined by our 5 milestone M1TM calculations that include chain lengths from 21 beads to 141 beads. We note the scaling exponents obtained from our brute force calculations, which include chain lengths from 21 to 101 beads agree within 1% of 5 milestone predictions. From Fig. 3.6, we see that indeed, the transit time and escape time scaling exponents exhibit excellent agreement, thus supporting the use of the escape time as a proxy for the transit time.

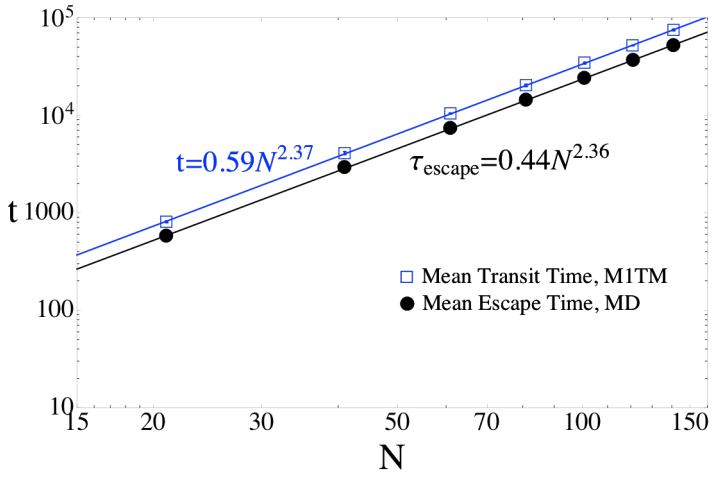


Figure 3.6: The mean transit time as computed by 5 milestone M1TM calculations and the mean escape time plotted as a function of chain length N. The straight lines are least squares fits to functions of the form $f(N) = bN^{\alpha}$.

As the MFPT is strictly greater than the MTT however, our results clearly cannot hold in the limit that $N \to \infty$. However, if we just consider chain lengths of 101 to 141 beads, the scaling exponents change to $\alpha = 2.20 \pm .12$ for the mean first passage time and $\alpha = 2.34 \pm .03$ for the MTT, which while within error of each other strongly suggest that longer chain lengths are necessary to resolve the asymptotic limits of these exponents. While it is reasonable to extend to longer chain lengths with calculations that use more milestones, it is not clear that the systematic error associated with such calculations will permit a very accurate determination of scaling exponents. We therefore defer an investigation of the long chain length scaling behavior to another work[78] that uses exact methods of calculation and considers the results in the context of the various theories of translocation.

3.6 Concluding Remarks

We have proposed a milestoning method for obtaining the probability distribution of transit times and the MTT of molecular processes. Like the case of MFPT, our result is given by a simple expression in closed form. Application of our method to the problem of polymer translocation yielded accurate estimates of the MTT for a good number of milestones. We also proposed to improve the efficiency of milestoning calculations by extracting statistics from the time reversed paths of sampled transitions. Application of the procedure to polymer translocation roughly doubled the observed number of sample transitions collected with a standard sampling procedure.

While our application was to a simple problem where a good reaction coordinate was available, we note that our approach to the transit time equation is general, as discussed in Appendices A and B, and may be applied to versions of milestoning that do not use a reaction coordinate as discussed in the appendices. Translocation of more detailed polymer models, such as proteins, has been explored in previous studies[79-81]. While these studies succeeded in shedding light on the thermodynimacs of large translocating polymers and the kinetics of small unstructured translocating polymers, they did not directly give translocation timescales for large / complex polymers owing to the infeasibility of applying brute force molecular dynamics simulations to the problem. With the transit time algorithm now available, milestoning represents a promising approach to obtaining the timescales of translocation for such systems

3.7 Appendix

3.7.1 General approach to the probability distribution of transit times

We now show how to calculate the transit time distribution using any type of milestoning—e.g. Directional Milestoning, Markovian milestoning with Voronoi tessellations, M1TM[23], MCM[82]. All of the above methods describe a trajectory by the sequence of "states" it passes through: $\{\alpha_1,\alpha_2,....,\alpha_n,...\}$. For example, such a state can be specified simply by the last milestone the trajectory crossed, or by the last 2 milestones it crossed[23]. The precise nature of those states is immaterial for the discussion below. In what follows, we assume states may be numbered one through *M* (the total number of states), and that Greek letters correspond to state numbers.

The reactants and products are no longer specified by single milestones but rather are collections of states. Reactive trajectories start in the reactants state and proceed to products, possibly through a series of intermediate states, without ever returning to the reactants (Fig. 3.7).

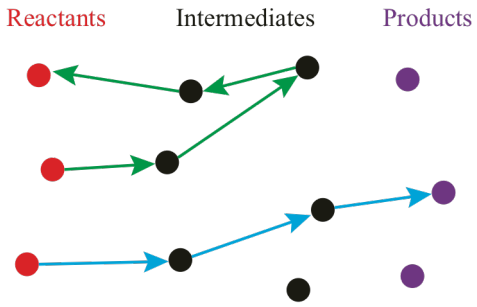


Figure 3.7: The circles represent (abstractly) the states in the milestoning state space. The green path is an example of an unreactive trajectory, which returns to the reactants before going to the products, while the blue path is an example of a reactive trajectory.

The probability distribution of transit times is given by a generalization of Eq. 3.5

$$P(t) = \sum_{\beta \notin \text{reactants}} \sum_{\alpha \in \text{reactants}} J_{\beta\alpha} \int_0^t u_\beta(t-t') P_{\beta\alpha}(t') dt' . \tag{3.21}$$

Eq. 3.21 in a weighted average over the transit time distributions corresponding to the different paths connecting the reactants to the products. Here $J_{\beta\alpha}$ is the probability that that *last* reactant state a trajectory visits is α, and that from there, the trajectory proceeds first to β and then onto the products without ever returning to reactants. The distribution $P_{\beta\alpha}(t)$ of transit times for the one step transition $\alpha \to \beta$ is obtained directly from the MD data, while $u_\beta(t)$ is the probability distribution of times *t* for the system, originating in β, to carry on to the products, conditional upon not first returning to the reactants, and is given by an expression similar to Eq. 3.6:

$$u_\beta(t-t') = \frac{\sum_{\gamma \in \text{products}} \hat{Q}_\gamma^\beta(t-t')}{p^\beta_{\text{products}}(\infty)}, \tag{3.22}$$

where $p^\beta_{\text{products}}(\infty)$ is the probability that a trajectory that has started in β will arrive in the products without first returning to the reactants, and $\hat{Q}_\gamma^\beta(t)$ is the probability flux into the product state γ. This flux is determined by Eq. 3.3, with the initial condition that the system starts in β and with absorbing boundary conditions at the reactant and product states. The splitting probability appearing in the denominator of Eq. 3.22 is given by Eqs. 3.7 and 3.8:

$$p^\beta_{\text{products}}(\infty) = \sum_{\gamma \in \text{products}} \mathbf{e}_\gamma^T \left[\mathbf{I} - \hat{\mathbf{T}}\right]^{-1} \mathbf{e}_\beta \ , \tag{3.23}$$

where summation accounts for all possible final states γ, $\mathbf{I}$ is the *M* by *M* identity matrix (recall *M* is the total number of states), $\mathbf{e}_\beta$ is the *M* dimensional unit column vector whose β^{th} component is 1 (all other components are 0), $\mathbf{e}_\gamma^T$ is the transpose of $\mathbf{e}_\gamma$ (defined analogously to $\mathbf{e}_\beta$), and $\hat{\mathbf{T}}$ is the modified transition probability matrix defined in accordance with the absorbing boundary conditions:

$$\hat{T}_{\mu\nu} \equiv \begin{cases} 0 & \text{if } \nu \in \text{ reactants or products} \\ T_{\mu\nu} & \text{otherwise} \end{cases}, \tag{3.24}$$

where $T_{\mu\nu}$ is the conditional probability a trajectory in ν will make its next transition to μ. Before moving on, we remark that Eqs. 3.21-3.23 are valid even when $\beta \in$ products

(i.e. transitions from the reactants directly to the products are permitted), as it can easily be verified that $\int_0^t u_\beta(t-t')P_{\beta\alpha}(t')dt' = P_{\beta\alpha}(t)$ in this case.

To obtain $J_{\beta\alpha}$ we write it as the product:

$$J_{\beta\alpha} = p^{\beta}_{\text{products}}(\infty)T_{\beta\alpha}\int_0^\infty Q_\alpha(t)dt\,, \tag{3.25}$$

where $\int_0^\infty Q_\alpha(t)dt$ is the volume of probability that passes through state α before the system arrives in the products, and $p^{\beta}_{\text{products}}(\infty)T_{\beta\alpha}$ represents the fraction of this volume that proceeds to the products through β without returning to the reactants.

Using Eqs. 3.7 and 3.8, we observe:

$$\int_0^\infty Q_\alpha(t)dt = \mathbf{e}_\alpha^T\left[\mathbf{I}-\mathbf{T}'\right]^{-1}\mathbf{p}(0)\,. \tag{3.26}$$

where $\mathbf{p}(0)$ is the vector of initial probabilities (i.e. its entries are $p_\alpha(0)$), and $\mathbf{T}'$ is the modified transition probability matrix whose entries are defined in accordance with the absorbing boundary conditions:

$$T'_{\mu\nu} \equiv \begin{cases} 0 & \text{if } \nu \in \text{ products} \\ T_{\mu\nu} & \text{otherwise} \end{cases}\,. \tag{3.27}$$

Using the definitions

$$\mathbf{e}_{\text{products}} \equiv \sum_{\gamma\in\text{products}} \mathbf{e}_\gamma\,, \tag{3.28a}$$

$$\mathbf{A} \equiv \left[\mathbf{I}-\hat{\mathbf{T}}\right]^{-1}, \tag{3.28b}$$

$$\mathbf{B} \equiv \left[\mathbf{I}-\mathbf{T}'\right]^{-1}, \tag{3.28c}$$

and Eqs. 3.23 and 3.26 to rewrite Eq. 3.25 as

$$J_{\beta\alpha} = \left(\mathbf{e}_{\text{products}}^T \mathbf{A} \mathbf{e}_\beta\right) T_{\beta\alpha} \left(\mathbf{e}_\alpha^T \mathbf{B} \mathbf{p}(0)\right). \tag{3.29}$$

Using Eqs. 3.22, 3.23, and 3.29 in Eq. 3.21, we conclude

$$P(t) = \sum_{\gamma \in \text{products}} \sum_{\beta \notin \text{reactants}} \sum_{\alpha \in \text{reactants}} T_{\beta\alpha} \left(\mathbf{e}_\alpha^T \mathbf{B} \mathbf{p}(0)\right) \int_0^t dt' P_{\beta\alpha}(t') \hat{Q}_\gamma^\beta (t - t') \tag{3.30}$$

Interestingly, we note that the splitting probabilities $p_{\text{products}}^\beta(\infty)$ have cancelled out of our final the expression. This does not contradict Eq. 3.12 however, because in the limit that milestoning with a reaction coordinate is used and the system is starts from milestone 1, the only conditional probability of the form $T_{\beta\alpha}$ is $T_{2,1} = 1$, and the volume of probability passing through milestone 1, $\mathbf{e}_1^T \mathbf{B} \mathbf{p}(0)$, is clearly the inverse of the splitting probability, thus Eq. 3.30 reduces to Eq. 12 in this limit. We further remark that while the splitting probabilities cancel out of Eq. 3.30, computing them is not a complete waste because as we shall see in the next subsection, they play into our expression for the MTT.

3.7.2 General formulation of the mean transit time

Similarly to Appendix 3.7.1, the MTT is given by an expression analogous to a result of section 3.3 (Eq. 3.19):

$$\langle t \rangle = \sum_{\beta \notin \text{reactantss}} \sum_{\alpha \in \text{reactant}} J_{\beta\alpha} \left(\langle \tau_\beta \rangle + \langle t_{\beta\alpha} \rangle \right), \tag{3.31}$$

where $\langle t_{\beta\alpha} \rangle$ is the MTT for the one step transition $\alpha \to \beta$, and $\langle \tau_\beta \rangle$ is mean the time for a trajectory initialized in β to terminate in a product state, conditional upon not

returning to any of the reactants; and $J_{\beta\alpha}$, given by Eq. 3.29 is the probability that the system takes this particular reaction path. We write

$$\left\langle \tau_\beta \right\rangle = \frac{\mathbf{e}_{\text{products}}^T \mathbf{A}\hat{\mathbf{L}}\mathbf{A}\mathbf{e}_\beta}{\mathbf{e}_{\text{products}}^T \mathbf{A}\mathbf{e}_\beta}, \tag{3.32}$$

where $\mathbf{A}$ is given be Eq. 3.28b, $\mathbf{L}$ is the matrix with elements $\hat{L}_{\mu\nu} \equiv \left\langle s_{\mu\nu} \right\rangle \hat{T}_{\mu\nu}$, $\left\langle s_{\mu\nu} \right\rangle$ is the mean lag time for the one step transition from ν to μ, and $\mathbf{e}_\beta$ and $\mathbf{e}_{\text{products}}$ are defined in Appendix 3.7.1. To obtain Eq. 3.32, we used

$$\left\langle \tau_\beta \right\rangle \equiv \int_0^\infty t u_\beta(t) dt = -\left. \frac{d\tilde{u}_\beta}{dz} \right|_{z=0}, \tag{3.33}$$

where $\tilde{u}_\beta(z)$ is the Laplace transform of $u_\beta(t)$, as defined by Eq. 3.22. Evaluation $-\left. \frac{d\tilde{u}_\beta}{dz} \right|_{z=0}$ is analogous to our expansion of Eq. 3.16. Our final result is therefore

$$\left\langle t \right\rangle = \sum_{\beta \notin \text{reactants}} \sum_{\alpha \in \text{reactants}} J_{\beta\alpha} \left(\frac{\mathbf{e}_{\text{products}}^T \mathbf{A}\hat{\mathbf{L}}\mathbf{A}\mathbf{e}_\beta}{\mathbf{e}_{\text{products}}^T \mathbf{A}\mathbf{e}_\beta} + \left\langle t_{\beta\alpha} \right\rangle \right). \tag{3.34}$$

3.7.3 Computing the probabilities, MFPTs, and MTTs of different reaction pathways

Aside from computing times scales of reactions with milestoning, we would also like to learn something about the sequence of conformational rearrangements that must take place in order for a reaction to occur. For example, if we have a system in which there is a large free energy barrier along the reaction coordinate, we can learn about the

conformational rearrangements that permit the reaction by looking at the conformational states the system assumes when it successfully crosses the free energy barrier. In section 2.4, we gave a numerical example where we computed the probability that each MCM state located at the free energy maximum is the *last* MCM state the system visits there before moving on to the products. Here, we detail how to compute the probability that an MCM state, α, is the final MCM state the system visits at a dividing surface before going on to the products. We also give expressions for the MFPT, and the MTT for trajectories that follow such a pathway. We conclude by noting that while our present treatment assumes both that a reaction coordinate is used to define the milestones of the system and that a single free energy barrier separates the products from the reactants, our approach can be generalized to systems that do not have a clear reaction coordinate and possess multiple free energy barriers.

The probability that α is the final state the system samples at the dividing surface, p_L^α, may be expressed as the product:

$$p_L^\alpha = f_{\text{products},\alpha} \int_0^\infty Q_\alpha(t)dt \,, \tag{3.35}$$

where the integral is the total volume of probability that passes through α, and can be simplified according to Eq. 3.8 to $\mathbf{e}_\alpha^T \left[\mathbf{I} - \mathbf{T}'\right]^{-1} \mathbf{p}(0)$. $f_{\text{products},\alpha}$ is the probability that the system, starting from α on the dividing surface, goes to the products without ever returning to the dividing surface once it leaves α. It is given by

$$f_{\text{products},\alpha} = \sum_{\beta \in \bar{\alpha}^+} \int_0^\infty dt Q''^{\beta}_{\text{productns}}(t) T_{\beta\alpha} = \sum_{\beta' \in \bar{\alpha}'^+} \left(\mathbf{e}^T_{\text{products}} \left[\mathbf{I} - \mathbf{T}''\right]^{-1} \mathbf{e}_\beta \right) T_{\beta\alpha} \tag{3.36}$$

where the summation is over all states in the set $\bar{\alpha}^+$, which are the states that lie between the dividing surface and products and can be reached from α via a single transition. $\int_0^\infty dtQ''^{\beta}_{\text{products}}(t)$ is the probability that the system, starting from β terminates in the products before returning to the dividing the surface. In its calculation, all states on the dividing surface and products are made absorbing. Eq. 3.8 was used to evaluate this integral in Eq. 3.36, where the elements of $\mathbf{T}''$ are defined according to the absorbing boundary conditions as follows:

$$T''_{\mu\nu} \equiv \begin{cases} 0 & \text{if } \nu \in \text{ dividing surface or products} \\ T_{\mu\nu} & \text{otherwise} \end{cases} . \tag{3.37}$$

Altogether then

$$p_L^{\alpha} = \sum_{\beta \in \bar{\alpha}^+} \left(\mathbf{e}_{\text{products}}^T \left[\mathbf{I} - \mathbf{T}'' \right]^{-1} \mathbf{e}_\beta \right) T_{\beta\alpha} \left(\mathbf{e}_\alpha^T \left[\mathbf{I} - \mathbf{T}' \right]^{-1} \mathbf{p}(0) \right). \tag{3.38}$$

The MFPT for paths that visit α on the dividing surface last, is given by

$$\tau_\alpha = \tau''_{\text{products},\alpha} + \frac{\int_0^\infty dtQ_\alpha(t)t}{\int_0^\infty dtQ_\alpha(t)} . \tag{3.39}$$

We explain the terms on the right hand side one at a time. The first term is the mean time for the system, initialized in α to go on to the products without returning to the dividing surface. It is given by

$$\tau''_{\text{products},\alpha} = \sum_{\beta \in \bar{\alpha}^+} p_{\text{productststs} \leftarrow \beta\alpha} \tau_{\text{products} \leftarrow \beta\alpha} \tag{3.40}$$

where $p_{\text{productsts}\leftarrow\beta\alpha}$ is the probability that the system, starting at α makes a transition directly to β and from there goes on to the products, conditional upon not returning to the dividing surface before arriving in the products. $\tau_{\text{products}\leftarrow\beta\alpha}$ is the time taken by such a path. Therefore ,

$$p_{\text{productsts}\leftarrow\beta\alpha} = \frac{\int_0^\infty dtQ''^{\beta}_{\text{products}}(t)T_{\beta\alpha}}{f_{\text{products},\alpha}} = \frac{\left(\mathbf{e}^T_{\text{products}}\left[\mathbf{I}-\mathbf{T}''\right]^{-1}\mathbf{e}_\beta\right)T_{\beta\alpha}}{f_{\text{products},\alpha}}. \qquad (3.41)$$

All terms in 3.41 have been defined earlier in this section, but we note this expression is sensible as the numerator is the absolute probability that the system reaches the products if it starts in α and proceeds through β, while the denominator is a rescaling factor applied so to make expression a conditional probability (the condition being that the system does not return to the dividing surface). The mean time for trajectories to go from α to the products through such a pathway is

$$\tau_{\text{productsts}\leftarrow\beta\alpha} = \frac{\int_0^\infty dtQ''^{\beta}_{\text{products}}(t)t}{\int_0^\infty dtQ''^{\beta}_{\text{products}}(t)} + \tau_{\beta\alpha} = \frac{\mathbf{e}^T_{\text{products}}\left[\mathbf{I}-\mathbf{T}''\right]^{-1}\mathbf{L}''\left[\mathbf{I}-\mathbf{T}''\right]^{-1}\mathbf{e}_\beta}{\mathbf{e}^T_{\text{products}}\left[\mathbf{I}-\mathbf{T}''\right]^{-1}\mathbf{e}_\beta} + \tau_{\beta\alpha} \quad (3.42)$$

where $\tau_{\beta\alpha}$ is the mean time for the trajectory to make the first leg of the journey from α to β, and the other term is the time it takes for the system to go from β onto the products. Upon evaluating the integrals with Eq. 3.8 and 3.16, we arrive at the right most expression of Eq. 3.42, where the only factor not defined is the matrix $\mathbf{L}''$, whose elements are given by $L''_{\mu\nu} = \langle s_{\mu\nu}\rangle T''_{\mu\nu}$, where we note $\langle s_{\mu\nu}\rangle$ was defined in section 3.7.2

as the mean lag time for a one step transition from ν to μ. Making the definition $\mathbf{D} \equiv \left[\mathbf{I} - \mathbf{T}''\right]^{-1}$ and subbing 3.41 and 3.42 into 3.40, we arrive at

$$\tau''_{\text{products},\alpha} = \sum_{\beta \in \bar{\alpha}^+} \frac{\left(\mathbf{e}^T_{\text{products}} \mathbf{D} \mathbf{e}_\beta\right) T_{\beta\alpha}}{f_{\text{products},\alpha}} \left(\frac{\mathbf{e}^T_{\text{products}} \mathbf{D} \mathbf{L}'' \mathbf{D} \mathbf{e}_\beta}{\mathbf{e}^T_{\text{products}} \mathbf{D} \mathbf{e}_\beta} + \tau_{\beta\alpha} \right) \tag{3.43}$$

With the first term in Eq. 3.39 taken care of, we observe the second term on the right hand side of Eq. 3.39 is the average time taken by the system to reach α for the last time. Our expression follows from recognizing that the probability of arriving at α for the last time at time *t*, is proportional to probability flux into α at time *t* multiplied by the probability that the system never returns to the dividing the surface after leaving: $Q_\alpha(t) f_{\text{products},\alpha}$. Upon multiplication by the normalization constant $\frac{1}{\int_0^\infty dt Q_\alpha(t) f_{\text{products},\alpha}}$, the probability distribution of final arrival times is simply $\frac{Q_\alpha(t)}{\int_0^\infty dt Q_\alpha(t)}$. Here, $Q_\alpha(t)$ is computed with just the product states made absorbing. Thus, using the same procedure as used to evaluate Eq. 3.16, we see the second term in Eq. 3.39 is given by $\frac{\mathbf{e}^T_\alpha \mathbf{B} \mathbf{L}' \mathbf{B} \mathbf{p}(0)}{\mathbf{e}^T_\alpha \mathbf{B} \mathbf{p}(0)}$, where the terms of the matrix $\mathbf{L}'$ are given by $L'_{\mu\nu} \equiv \langle s_{\mu\nu} \rangle T'_{\mu\nu}$, and we recall that $\mathbf{B} \equiv \left[\mathbf{I} - \mathbf{T}'\right]^{-1}$ according to definition 3.28c. Thus MFPT, conditional upon α being the final state the system visits on the dividing surface, is:

$$\tau_\alpha = \frac{\mathbf{e}^T_\alpha \mathbf{B} \mathbf{L}' \mathbf{B} \mathbf{p}(0)}{\mathbf{e}^T_\alpha \mathbf{B} \mathbf{p}(0)} + \tau''_{\text{products},\alpha} \tag{3.44}$$

with $\tau''_{\text{products},\alpha}$ given by Eq. 3.43.

To obtain the MTT for trajectories that pass through α last on the dividing surface, we must perform a weighted average over all reactive trajectories that begin in the reactants:

$$t_\alpha = \frac{\sum_{\mu \notin \text{reactants}} \sum_{\nu \in \text{reactants}} C_{\alpha\mu} J_{\mu\nu} \left(t_{\mu\nu} + \hat{\tau}_{\alpha,\mu} + \tau''_\alpha \right)}{\sum_{\mu \notin \text{reactants}} \sum_{\nu \in \text{reactants}} C_{\alpha\mu} J_{\mu\nu}} . \tag{3.45}$$

Here, $J_{\mu\nu}$, given by Eq. 3.25, is the absolute probability that the last state the system visits in the reactants before going on to the products is ν and that the system transitions directly to μ after its final visit to ν. $C_{\alpha\mu}$ is the conditional probability that the last state the system visits on the dividing surface before going on to the products is α, provided it starts in μ and does not return to the reactants before going on to the products. The denominator of Eq. 3.45 ensures normalization. $C_{\alpha\mu}$ can be expressed as

$$C_{\alpha\mu} = \frac{f_{\text{products},\alpha} \int_0^\infty dt \hat{Q}^\mu_\alpha(t)}{\int_0^\infty dt \hat{Q}^\mu_{\text{products}}(t)} = \frac{f_{\text{products},\alpha} \mathbf{e}_\alpha^T \left[\mathbf{I} - \hat{\mathbf{T}} \right] \mathbf{e}_\mu}{\mathbf{e}_{\text{products}}^T \left[\mathbf{I} - \hat{\mathbf{T}} \right] \mathbf{e}_\mu} , \tag{3.46}$$

where numerator gives the absolute probability that the system reaches the products with α being the last state at the dividing surface the system visits, and the conditions that the system starts in μ and does not return to the reactants. Note that $\hat{Q}^\mu_\alpha(t)$ and $\hat{\mathbf{T}}$ were defined in the previous 2 appendices with absorbing boundary conditions at the reactants and products. The denominator of Eq. 3.46 is the absolute probability that the system reaches the products before returning to the reactants, and has the effect of making the

expression in Eq. 3.45 the appropriate conditional probability. $t_{\mu\nu}$ is the transit time for the system to go from ν to μ (measured directly from the MD data), $\hat{\tau}_{\alpha,\mu}$ is the mean time for a trajectory to reach α for the final time conditional upon starting at μ and not returning to reactants, and $\tau''_{\text{products},\alpha}$ is the time (as defined in Eq. 3.40 and expressed in Eq. 3.44) for the system to go on from α to the products without returning to the dividing surface. To obtain $\hat{\tau}_{\alpha,\mu}$ we note similarly to what we did in obtaining the 2nd term of Eq. 3.39, that the probability distribution of *last* arrival times on α is proportional to the probability flux into α multiplied by the probability that the system never returns to the reactants: $f_{\text{products},\alpha}\hat{Q}^{\mu}_{\alpha}(t)$. Thus $\dfrac{\hat{Q}^{\mu}_{\alpha}(t)}{\int_0^{\infty} dt\hat{Q}^{\mu}_{\alpha}(t)}$ is the probability distribution of last arrival times on α, conditional upon starting in μ and not returning to the reactants. Therefore,

$$\hat{\tau}_{\alpha,\mu} = \frac{\int_0^{\infty} dt\hat{Q}^{\mu}_{\alpha}(t)t}{\int_0^{\infty} dt\hat{Q}^{\mu}_{\alpha}(t)} = \frac{\mathbf{e}_a^T\left[\mathbf{I}-\hat{\mathbf{T}}\right]^{-1}\hat{\mathbf{L}}\left[\mathbf{I}-\hat{\mathbf{T}}\right]^{-1}\mathbf{e}_{\mu}}{\mathbf{e}_a^T\left[\mathbf{I}-\hat{\mathbf{T}}\right]^{-1}\mathbf{e}_{\mu}}. \tag{3.47}$$

Using the definitions 3.28b and 3.28c, $\mathbf{A} \equiv \left[\mathbf{I}-\hat{\mathbf{T}}\right]^{-1}$ and $\mathbf{B} \equiv \left[\mathbf{I}-\mathbf{T}'\right]^{-1}$, the MTT, conditional upon α being the final state the system visits on the dividing surface, is:

$$t_{\alpha} = \frac{\sum_{\nu\notin\text{reactants}}\sum_{\mu\in\text{reactants}} f_{\text{products},\alpha}\left(\mathbf{e}_{\alpha}^T\mathbf{A}\mathbf{e}_{\mu}\right)T_{\mu\nu}\left(\mathbf{e}_{\nu}^T\mathbf{B}\mathbf{p}(0)\right)\left(\tau''_{\text{products},\alpha} + \dfrac{\mathbf{e}_{\alpha}\mathbf{A}\hat{\mathbf{L}}\mathbf{A}\mathbf{e}_{\mu}}{\mathbf{e}_{\alpha}^T\mathbf{A}\mathbf{e}_{\mu}} + t_{\mu\nu}\right)}{\sum_{\nu\notin\text{reactants}}\sum_{\mu\in\text{reactants}} f_{\text{products},\alpha}\left(\mathbf{e}_{\alpha}^T\mathbf{A}\mathbf{e}_{\mu}\right)T_{\mu\nu}\left(\mathbf{e}_{\nu}^T\mathbf{B}\mathbf{p}(0)\right)} \tag{3.48}$$

where Eq. 3.36 gives $f_{\text{products},\alpha}$, and Eq. 3.43 gives $\tau''_{\text{products},\alpha}$.

We conclude this appendix by noting that it is possible to extend the approach taken in this section to a much more general problem, which can be stated as follows. Given *n* groups of states $\{G_1, G_2, ..., G_n\}$, what is the probability of seeing a path $\{..., \alpha_1, ..., \alpha_2, ..., \alpha_n\}$ in which for each *i*, the last visit to α_i precedes the last visit to α_{i+1} and α_i is the last state visited in G_i before going on to the products? What are the MFPT and MMT for such a path? Expressions for these quantities can be given in closed form, but for economy of space we, do not perform the derivation here.

3.7.4 Models of the polymer and pore

3.7.4.1 Polymer model

Our polymer model consists of *N* beads of mass m connected by *N*-1 springs. The pairwise interaction potential between adjacent beads is:

$$V_{bond}(r) = k_{bond}(r - \sigma)^2 / 2 \tag{3.49}$$

The non-bonded beads interact via a repulsive Leonard-Jones potential representing excluded volume effects:

$$V_{non-bonded}(r) = 4\varepsilon\left[(\sigma / r)^{12} - (\sigma / r)^6\right]\theta(2^{1/6}\sigma - r) \tag{3.50}$$

Here σ is the equilibrium bond distance, ε is the characteristic energy scale, $k_{bond} = 100\varepsilon / \sigma^2$, and θ is the Heaviside step function that truncates the attractive portion of the Lennard Jones potential.

The dynamics of each bead is governed by the Langevin equation:

$$m\ddot{r}_i(t) = -\frac{\partial V}{\partial r_i} - \xi \dot{r}_i(t) + f(t) \tag{3.51}$$

where $r_i(t)$ is the position of the bead, V is the total interaction potential, ξ is the friction coefficient, and $f(t)$ is a random force satisfying the fluctuation dissipation theorem. Our data is reported in reduced units, where the units of length and energy are σ and ε. In our simulations, the unit of time is $t = \sqrt{m\sigma^2 / \varepsilon}$, the temperature $T = 1.0\varepsilon / k_B$, and our friction coefficient is $\xi = 2.0(\sigma^2 / m\varepsilon)^{-1/2}$.

3.7.4.2 The Pore

The interaction between the polymer and the cylindrical pore is given by

$$V_{pore} = \sum_{i=1}^{N+1} \frac{a}{1+e^{-z_i'/l}} \left(1 - \frac{1}{1+\left(\frac{x_i^2 + y_i^2}{r_{pore}} \right)} \right), \tag{3.52}$$

where $a = 20\varepsilon$, $l = \sigma/20$, and $z' = \begin{cases} z_i & \text{if } z_i < d/2 \\ d - z_i & \text{if } z_i \geq d/2 \end{cases}$.

The length of the pore is roughly given by $d = 1.5\sigma$, and the radius is $r = r_{pore} + \delta r + r_{monomer}$. $r_{pore} + \delta r$ represents the effective pore radius the monomer center of mass may move within, and $r_{monomer} = \sigma/2$ is the monomer radius. δr can be found by solving the equation[36]

$$\int_0^{\infty} 2\pi r e^{-V_{pore}(r,z')/K_BT}\, dr = \int_0^{r_{pore}+\delta r} 2\pi r\, dr \tag{3.53}$$

To make our study consistent with previous literature [65-71], we desire a pore radius of roughly σ. Taking $r_{pore} = 0.6\sigma$, $z' = d/2$, and solving Eq. 3.39 for δr, we obtain $r = 1.00\sigma$. We note that the value of δr is not heavily dependent on the value of z' as long as it is taken between 0 and $d/2$ (e.g. at $z' = 0$, the effective pore radius is $r = 1.02\sigma$).

3.7.4.3 The reflecting boundary

The reflecting boundary as seen by monomer 1 is taken as the $z = d/2$ plane, and we impose it by including the potential

$$V_b = \theta\left(d/2 - z_1\right)\frac{k\left(z_1 - d/2\right)^2}{2}\ , \tag{3.54}$$

where θ is a heavy side step function and $k = 1000\varepsilon/\sigma^2$.

Chapter 4. Using simple polymer models to explore the role of internal friction in the dynamics of unfolded proteins

4.1 Abstract

Recent experiments showed that the reconfiguration dynamics of unfolded proteins are often adequately described by simple polymer models. In particular, the Rouse model with internal friction (RIF) captures internal friction effects as observed in single-molecule fluorescence correlation spectroscopy (FCS) studies of a number of proteins. Here we use RIF, and its non-free draining analog, Zimm model with internal friction (ZIF), to explore the effect of internal friction on the rate with which intramolecular contacts can be formed within the unfolded chain. Unlike the reconfiguration times inferred from FCS experiments, which depend linearly on the solvent viscosity, the first passage times to form intramolecular contacts are shown to display a more complex viscosity dependence. We further describe scaling relationships obeyed by contact formation times in the limits of high and low internal friction. Our findings provide experimentally testable predictions that can serve as a framework for the analysis of future studies of contact formation in proteins.

This chapter was co-written by Ryan R. Cheng, Alexander T. Hawk (co-first authors), and Dmitrii E. Makarov. Ryan made contributions to the equations presented herein, performed all simulations, and assisted in establishing comparisons with experiments. Alexander developed the framework for the Zimm with internal friction model and made contributions to the general discussion of internal friction. Dr. Makarov initiated the project, made contributions to the general theoretical framework presented, and provided interpretations for our data.

4.2 Introduction

The timescales at which unfolded proteins diffusively undergo conformational rearrangement are intimately related to the "speed limit" of protein folding[83-87] and have consequently attracted considerable attention[88-97]. Rationalization of the various experiments that probe those timescales requires models that adequately capture the dynamics in the unfolded state. While potentially neglecting certain structural details[98, 99], simple polymer models, such as the exclude-volume random coil, are often capable of accounting for experimentally measurable properties of equilibrium conformational ensembles of denatured proteins with remarkable accuracy[100-104]. In contrast, the ability of polymer theory to account for the *dynamics* of unfolded proteins remains a largely unexplored issue. In fact, with only a few exceptions[94, 105, 106], most of the existing experimental data has so far been interpreted in terms of even simpler, one-dimensional models, where complex polymer dynamics is viewed as diffusion along an appropriately chosen reaction coordinate[107, 108]. Such greatly simplified models, however, cannot capture multidimensional properties as displayed, for example, by the dependence of the characteristic reconfiguration times on the location of the probes used to measure them within the polypeptide chain[97]. Moreover, they usually neglect essential non-Markovian effects in the dynamics of the one-dimensional reaction coordinates[108-112].

A further challenge to theoretical models of protein dynamics lies in disentangling the solvent and internal friction effects. It has been long recognized[113-120] that, in addition to solvent friction, other dissipative mechanisms (i.e., internal friction) may play a

significant role in the dynamics and folding of proteins. Internal friction reflects the intrinsic resistance of a polymer to changes in its conformation and stems from the "roughness" of the underlying energy landscape[18, 121]. Possible mechanisms that contribute to internal friction include dihedral angle rotational barriers[18, 122], hydrogen bonding[123] and intrachain collisions[116]. While observations of internal friction effects have been reported[94, 105, 106], our understanding of these effects is limited, particularly because they were commonly interpreted in terms of phenomenological models that lack rigorous justification and do not provide a direct link between observations and the microscopic mechanisms of friction. As a step toward a more quantitative picture, a recent single-molecule FRET study of the reconfiguration dynamics of unfolded *T. maritime* cold shock protein and of intrinsically disordered proteins[94] provided a more quantitative account of internal friction in terms of a well established model called Rouse with internal friction (RIF) [15-19], suggesting that polymer theory may provide a sufficient framework to describe the multidimensional dynamics of the unfolded state. Of course, extensions of that study to other proteins and other types of experimental measurements would be necessary to further validate this view. In anticipation of such future studies, this paper explores experimentally verifiable predictions of RIF and related models.

The idea of applying RIF to proteins is not new: It has, for example, been discussed in the context of protein folding[18], applied to strongly stretched biopolymers[19, 124], and used to demonstrate the complex interplay between the solvent- and internal friction in the dynamics of polypeptides[125]. Given its simplicity that often affords analytic

solutions, RIF provides an attractive alternative to the computationally costly molecular dynamics simulations, although the latter would still be required to elucidate the molecular mechanisms of internal friction.

RIF, however, has several significant limitations. First, it ignores hydrodynamic interactions within the polymer chain. To address this drawback, here we also develop a non-free draining version of RIF called Zimm with internal friction (ZIF), which was qualitatively discussed earlier in the context of protein folding[18]. While incorporating internal friction in a similar manner, the more realistic ZIF builds upon the Zimm model[126], which accounts for hydrodynamic effects. Second, the dimensions of unfolded proteins may depend on the denaturant concentration or temperature[127-129], with low- or zero-denaturant cases favoring more compact, collapsed conformations. Although coil-globule transition theory was shown to successfully account for such changes in equilibrium dimensions and conformational statistics[104, 128, 130], it is an equilibrium theory that makes no predictions regarding the dynamics. Here, we employ a simple confining harmonic potential to mimic protein collapse, motivated by the earlier evidence[94] that such modification of the potential adequately accounts for the collapse-induced changes in the conformational statistics of more realistic polymer models. Finally, excluded volume effects are ignored by RIF and ZIF. While potentially important, they are not considered here, as our earlier study[105] suggested that their consequences are often too subtle to be resolved in experimental studies.

Using RIF and ZIF, we attempt to mimic typical experiments that probe the

dynamics of the unfolded state with the goal of establishing how internal friction manifests itself in those experiments. Specifically, we examine two well-established experimental methodologies. The first one uses single-molecule nanosecond fluorescence correlation spectroscopy (nsFCS), which employs fluorescence resonance energy transfer (FRET) between a pair of dyes attached at different locations along the peptide backbone. Fluctuations in the fluorescence intensities of the donor and acceptor are related to the fluctuations in their mutual distance and the characteristic timescale associated with these fluctuations is taken to be the reconfiguration time of the protein[92, 96, 97, 131]. Furthermore, by varying the positions of the donor and acceptor, one could probe the spectrum of reconfiguration times of an unfolded protein[106]. We note that RIF has already been applied to nsFCS data[94]; here the key findings of that work will be recapitulated and extended to include the ZIF case.

The second type of experiment involves the use of a fluorescent probe and a quencher located at different positions along a peptide chain. In this case, the fluorescence lifetime of the probe is reduced by its collisions with the quencher, thus providing a method to measure the time it takes for the probe and quencher to diffusively come into a close contact and to form a loop within the chain[84, 85, 89]. The timescales inferred from the two types of experiments are not the same, although they are related to each other. Indeed, fluorescence quenching measurements probe the time it takes the quencher and the probe to come into close proximity; consequently, they yield longer timescales than those exhibited by fluctuations of the long-range energy transfer

employed in nsFCS measurements. It is then plausible that internal friction affects the two timescales differently. Indeed, in the following sections we find this to be the case. Our study, then, provides a theoretical framework for analyzing future fluorescence quenching experiments aimed at measuring internal friction effects: If they turn out to be consistent with the RIF/ZIF predictions described here, this will provide further experimental support of RIF/ZIF as a viable model of protein dynamics in the unfolded state.

This chapter is organized as follows: Section 4.3 introduces the models used in our study and provides simulation details. Section 4.4 describes the effect of internal friction, as predicted by ZIF and RIF, on quantities measured in nsFCS experiments. Section 4.5 discusses the effect of internal friction on the timescales of loop closure and provides a comparison of polymer theory predictions with experimental loop closure times of a peptide. Finally, section 4.6 concludes with closing remarks.

4.3 Model details

All of the polymer models considered here consist of N+1 beads with coordinates $\mathbf{r} = (\mathbf{r}_0, \mathbf{r}_1, ..., \mathbf{r}_N)$. The equations describing the Brownian dynamics of the beads can be written in the general form[126]:

$$\frac{d\mathbf{r}}{dt} = \mathbf{HF} \tag{4.1}$$

Here $\mathbf{F} = (\mathbf{F}_0, \mathbf{F}_1, ..., \mathbf{F}_N)$ represents the sum of forces on each bead and **H** is the mobility tensor, with components $\mathbf{H}_{nm} \equiv \mathbf{H}(\mathbf{r}_n - \mathbf{r}_m)$. Note that $\mathbf{H}_{nm}$ is a 3 by 3 matrix rather than a scalar.

4.3.1. Rouse and Rouse with internal friction (RIF)

The Rouse model[16, 126] is a coarse grained model, which considers polymers at a spatial resolution coarser than their Kuhn length. It consists of a chain of beads connected by linear springs and neglects excluded-volume interactions among the beads. It further neglects hydrodynamic interactions so that $\mathbf{H}_{nm} = \mathbf{I}\delta_{nm} / \xi_s$, where ξ_s is the solvent friction coefficient, and **I** is the 3×3 identity matrix. Consequently, Eq. 4.1 simplifies to

$$\xi_s \frac{d\mathbf{r}}{dt} = \kappa \mathbf{k}\mathbf{r} + \mathbf{f} \tag{4.2}$$

where $\kappa = 3k_B T / b^2$ is the stiffness of the connecting springs, *b* is the Kuhn length, $\mathbf{f} = (\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_N)$ are the random forces on each bead and **k** is the tri-diagonal connectivity matrix:

$$\mathbf{k} = \begin{pmatrix} -1 & 1 & 0 & 0 & ... \\ 1 & -2 & 1 & 0 & ... \\ 0 & 1 & -2 & 1 & ... \\ 0 & 0 & 1 & -2 & ... \\ ... & ... & ... & ... & ... \end{pmatrix} \tag{4.3}$$

The microscopic dynamics occurring at spatial scales shorter than the Kuhn length (i.e. hindered dihedral rotations) results in internal friction, which is included in RIF in a semi-phenomenological manner[15-17, 19] through the introduction of a frictional

drag term that resists the relative extension between connected beads. In this regard, the internal friction coefficient is the result of polymer coarse graining in much the same way as solvent friction results from not explicitly considering the solvent molecules in the Rouse model. The resulting equation of motion for the RIF includes both solvent and internal friction effects and is given by

$$\xi_s \frac{d\mathbf{r}}{dt} = \kappa \mathbf{k}\mathbf{r} + \xi_i \frac{d}{dt}\mathbf{k}\mathbf{r} + \mathbf{f}^* \tag{4.4}$$

where ξ_i is the internal friction coefficient and $\mathbf{f}^*$ is the appropriate random force vector[19].

The set of equations described by Eq. 4.4 can be decoupled by making a coordinate transformation to the eigen-basis of the connectivity matrix—i.e., normal mode decomposition. The eigenvectors of **k** are given by

$$u_{pn} = \frac{1}{N+1}\cos\left(\frac{p\pi}{N+1}\left(n+\frac{1}{2}\right)\right) \tag{4.5}$$

and have the associated eigenvalues

$$\lambda_p = 4\sin^2\left(\frac{p\pi}{2(N+1)}\right) \tag{4.6}$$

where $p = 0,1,...,N$ denotes the mode index and $n = 0,1,...,N$, again, enumerates the monomers. The dynamics of the chain can thus be written as a superposition of independent eigenmodes:

$$\mathbf{r}(t) = \sum_{p=0}^{N} \mathbf{X}_p(t)\mathbf{u}_p \tag{4.7}$$

In the basis of Eq. 4.5, Eq. 4.4 results in N+1 independent overdamped Langevin equations:

$$\xi_p^{(RIF)} \frac{d\mathbf{X}_p}{dt} = -k_p \mathbf{X}_p + \left(2\xi_p^{(RIF)} k_B T\right)^{1/2} \mathbf{a}_p(t) \tag{4.8}$$

For the p-th mode, $\xi_p^{(RIF)} = \xi_s + \xi_i \lambda_p$ is the effective friction coefficient, $k_p = \kappa \lambda_p$ is the effective spring coefficient (identical to that of the Rouse model), and $\mathbf{a}_p = (a_{p,x}, a_{p,y}, a_{p,z})$ is a white Gaussian noise process characterized by the moments $\langle \mathbf{a}_p(t) \rangle = 0$ and $\langle a_{p\alpha}(t) a_{q\beta}(t') \rangle = \delta_{pq}\delta_{\alpha\beta}\delta(t-t')$, where p and q denote the mode index while α and β denote the x, y, or z directions.

The Langevin equation for each RIF mode describes an overdamped harmonic oscillator with a characteristic relaxation time

$$\tau_p^{(RIF)} = \frac{\xi_p^{(RIF)}}{k_p} \cong \frac{\tau_{Rouse}}{p^2} + \tau_i \tag{4.9}$$

where $\tau_{Rouse} = \xi_s N^2 b^2 / (3\pi^2 k_B T)$ is the slowest relaxation time of the Rouse model and $\tau_i = \xi_i / \kappa$ is the relaxation time due to internal friction. In the absence of internal friction (i.e., $\xi_i = 0$), the equation of motion for each independent mode and the spectrum of relaxation times of the Rouse model are recovered from Eq. 4.8 and Eq. 4.9, respectively. Because τ_i is independent of the chain length N, the effect of internal friction on the slowest relaxation time of RIF becomes negligible for polymers of sufficient length,

where $\tau_1^{RIF} \approx \tau_{Rouse}$, a statement known as the Kuhn theorem[16]. However internal friction effects may still be significant for the measurements that probe faster relaxation modes.

4.3.2 Zimm with internal friction (ZIF)

The Zimm model[126] accounts for hydrodynamic interactions using the pre-averaging approximation, where $\mathbf{H}_{nm}$ is replaced by its equilibrium average:

$$\mathbf{H}_{nm} \to \langle \mathbf{H}_{nm} \rangle = \frac{\mathbf{I}}{2\pi^2 \xi_s |n-m|^{1/2}} \tag{4.10}$$

Treating the internal friction contribution in the same way as above (cf. Eq. 4.4), Eq. 4.1 then becomes

$$\frac{d\mathbf{r}}{dt} = \langle \mathbf{H} \rangle \left(k_0 \mathbf{kr} + \xi_i \frac{d}{dt} \mathbf{kr} + \mathbf{f}^* \right). \tag{4.11}$$

For $N \gg 1$, $\langle \mathbf{H} \rangle$ becomes diagonal in the eigenbasis of the connectivity matrix. Therefore, similarly to Eq. 4.4, Eq. 4.11 reduces to N+1 independent overdamped Langevin equations:

$$\xi_p^{(ZIF)} \frac{d\mathbf{X}_p}{dt} = -k_p \mathbf{X}_p + \left(2\xi_p^{(ZIF)} k_B T \right)^{1/2} \mathbf{a}(t) \tag{4.12a}$$

$$\begin{aligned} \xi_p^{(ZIF)} &= \xi_s \left(\frac{\pi p}{3(N+1)} \right)^{1/2} + \xi_i \lambda_p \\ \xi_0^{(ZIF)} &= \xi_0^{(Zimm)} = \xi_s \left(\frac{3\pi}{32N} \right)^{1/2} \end{aligned} \tag{4.12b}$$

For the p-th mode, $\xi_p^{(ZIF)}$ is the effective friction coefficient, $k_p = \kappa\lambda_p$ is the effective spring constant, and $\mathbf{a}_p$ is the same white Gaussian noise process as in Eq. 4.8. Likewise, the relaxation time of each ZIF mode is written as

$$\tau_p^{(ZIF)} = \frac{\xi_p^{(ZIF)}}{k_p} \cong \frac{\tau_{Zimm}}{p^{3/2}} + \tau_i \tag{4.13}$$

where $\tau_{Zimm} = \xi_s (N+1)^{3/2} b^2 / (3\sqrt{3} k_B T \pi^{3/2})$ is the slowest relaxation time of the Zimm chain and $\tau_i = \xi_i / \kappa$ is the relaxation time due to internal friction. In the absence of internal friction (i.e., $\xi_i = 0$), the equation of motion for each independent mode and the spectrum of relaxation times of the Zimm model are recovered from Eq. 4.12 and Eq. 4.13, respectively.

4.3.3 Simulation details

Time evolution of RIF and ZIF was computed from the time evolution of Eq. 4.8 and Eq. 4.12. The bead positions, as a function of time, were then recovered through the linear transformation of Eq. 4.7. The end-to-end loop formation time, τ_{loop}, was computed from the time dependence of the end-to-end distance, $R(t) = \left| \mathbf{r}_N(t) - \mathbf{r}_0(t) \right|$, as a mean first passage time for the ends of a polymer chain to diffusively come within a distance R_c of one another (i.e., $R < R_c$) when starting from the equilibrium ensemble of conformations with $R > R_c$ [105].

4.4 Reconfiguration times

In earlier studies[94, 106], we have used the Rouse and RIF models to account for FRET-derived reconfiguration times in unfolded proteins. Here we recapitulate those results and extend them to the ZIF case.

4.4.1 Definition of reconfiguration time.

Consider the autocorrelation function of the distance vector between a pair of monomers *n* and *m*:

$$\phi_{nm}(t) = \left\langle \left(\mathbf{r}_n(t) - \mathbf{r}_m(t)\right) \bullet \left(\mathbf{r}_n(0) - \mathbf{r}_m(0)\right) \right\rangle \tag{4.14}$$

A reconfiguration time, τ_{nm}, can be defined from the relaxation of $\phi_{nm}(t)$ [94, 106]:

$$\tau_{nm} = \int_0^\infty dt \phi_{nm}(t) / \phi_{nm}(0) \tag{4.15}$$

If $\phi_{nm}(t)$ exhibits exponential decay, i.e., $\phi_{nm}(t) = \phi_{nm}(0)\exp(-t/\tau_{nm})$, then the definition of Eq. 4.15 recovers the decay time. While a single timescale cannot be uniquely defined if $\phi_{nm}(t)$ is nonexponential, Eq. 4.15 still provides a reasonable measure of a reconfiguration time[132, 133].

For RIF or ZIF, Eqs. 4.14 and 4.15 can be evaluated analytically. Substituting Eq. 4.7 into Eq. 4.14 results in

$$\phi_{nm}(t) = \sum_{p=1}^{N}\sum_{q=1}^{N} \left\langle \mathbf{X}_p(t) \cdot \mathbf{X}_q(0) \right\rangle \left(u_{pn} - u_{pm}\right)\left(u_{qn} - u_{qm}\right) \tag{4.16}$$

where *p* and *q* denote the mode index. Note that the translational mode, i.e. $p,q = 0$, does

not contribute to the summations since $u_{0n} - u_{0m} = 0$. The correlation functions appearing in Eq. 4.16 are readily evaluated given that the variables $\mathbf{X}_p(t)$ represent a set of statistically independent overdamped harmonic oscillators (cf. Eqs. 4.8 and 4.12):

$$\left\langle X_{p\alpha}(t) \cdot X_{q\beta}(0) \right\rangle = \left(k_B T / k_p \right) \delta_{pq} \delta_{\alpha\beta} \exp\left(-t \cdot k_p / \xi_p \right) \tag{4.17}$$

where α and β denote the *x*, *y*, or *z* directions and δ is the Kronecker delta. Applying Eq. 4.17 to Eq. 4.16 yields

$$\phi_{nm}(t) = 3k_B T \sum_{p=1}^{N} k_p^{-1} \exp\left(-t \cdot k_p / \xi_p \right) \left(u_{pn} - u_{pm} \right)^2 \tag{4.18}$$

The reconfiguration time can then be obtained by substituting Eq. 4.18 into Eq. 4.15:

$$\tau_{nm} = \frac{\xi_0}{\kappa} \frac{\sum_{p=1}^{N} (\lambda_p)^{-2} (u_{pn} - u_{pm})^2}{\sum_{p=1}^{N} (\lambda_p)^{-1} (u_{pn} - u_{pm})^2} + \tau_i \tag{4.19}$$

where $\xi_0 = \xi_0^{(Rouse)} = \xi_s$ for RIF and $\xi_0 = \xi_0^{(Zimm)}$ is given by Eq. 4.12 for ZIF. We will refer to the reconfiguration time given by Eq. 4.19 as $\tau_{nm}^{(RIF)}$ and $\tau_{nm}^{(ZIF)}$ to differentiate between the two models.

It should be noted that the reconfiguration time defined here is related to the *vector*, $\mathbf{r}_n - \mathbf{r}_m$, that represents the spatial separation between two monomers, while the reconfiguration time in nsFCS studies is a function of the fluctuations of the *absolute distance*, $\left| \mathbf{r}_n - \mathbf{r}_m \right|$. The choice of the former leads to a great mathematical simplification and affords simple analytic solutions. It may, however, be viewed as unphysical because,

even if the distance between the monomers is not changing, the distance vector may relax through rotations, which, on the other hand, do not affect the FRET signal. In defense of this simplification, we previously demonstrated[94, 106] that the prediction of Eq. 4.19 is in a good agreement with simulations of FRET in more realistic polymer models. If so desired, of course, the inter-monomer distance autocorrelation time is equally easy to estimate numerically using polymer trajectories as described in Section 4.3.3.

4.4.2 Internal friction has additive effect on reconfiguration times

One of the most common methods of quantifying internal friction is to vary solvent viscosity, η_s. Provided that the timescale of interest depends linearly on η_s, the intercept obtained for $\eta_s \to 0$ then provides a measure of solvent-independent, internal friction. This approach has been used in studies of conformational relaxation of myoglobin[115], kinetics of protein folding[113, 116-119, 134], and reconfiguration dynamics of intrinsically disordered or chemically denatured proteins[94].

Both RIF and ZIF, indeed, predict a linear viscosity dependence of the reconfiguration time, which can be extrapolated to obtain the internal friction time constant τ_i. This can be seen from Eq. 4.19, which is of the form

$$\tau_{nm}^{(RIF,ZIF)} = \tau_{nm}^{(Rouse,Zimm)} + \tau_i \tag{4.20}$$

where the reconfiguration time of the Rouse/Zimm model[106] is recovered if $\tau_i \to 0$. Since $\tau_{nm}^{(Rouse)}$ and $\tau_{nm}^{(Zimm)}$ are both proportional to the solvent viscosity and assuming that τ_i is independent of the viscosity, Eq. 4.20 gives a linear viscosity dependence (Fig. 4.1).

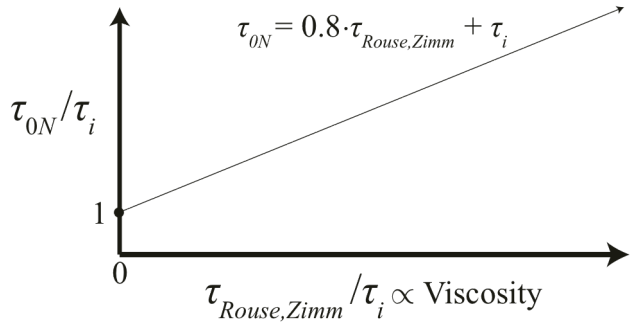


Figure 4.1: This figure illustrates that the reconfiguration time exhibits additive contribution of internal friction and linear dependence on solvent viscosity. Both RIF and ZIF predict a linear viscosity dependence of the reconfiguration times, with a finite intercept of τ_i. When the end-to-end reconfiguration time is rescaled by τ_i and is plotted as function of the slowest relaxation time (of Rouse or Zimm) rescaled by τ_i, the dependence is a straight line with a slope of ≈ 0.8 and intercept of 1 (cf. Eq. 4.21). Moreover, all reconfiguration times (for any pair of monomers *n* and *m*) have the same intercept, τ_i. That is, in an internal friction dominated regime the *n* and *m* dependence of τ_{nm} becomes weak, a trend, indeed, observed experimentally[94]. For the end-to-end dynamics (i.e., $n = 0$ and $m = N$) in the long chain limit ($N \gg 1$), Eq. 4.20 approximately gives

$$\tau_{0N}^{(RIF,ZIF)} \approx 0.8\tau_{Rouse,Zimm} + \tau_i \tag{4.21}$$

where $\tau_{Rouse,Zimm}$ is the relaxation time of the slowest Rouse (or Zimm) mode given in Section 4.3. Interestingly, the factor of ~0.8 is coincidentally the same for both RIF and ZIF.

We note that, although RIF and ZIF predict different chain length dependences for the reconfiguration times, in practice both ZIF and RIF fits of experimental data may be of comparable quality because of the relatively short chain lengths typical of single-

domain proteins employed in nsFCS studies. Indeed, both ZIF (assumed empirically in ref. [94]) and RIF yield essentially the same estimate for τ_i for the proteins studied in ref. [94].

Finally, it should be noted that both the RIF and ZIF have the limitation of applying internal friction to all non-translational normal modes of the system. While the equations of motion are greatly simplified as a consequence, this treatment is not entirely correct, as it results in the unphysical prediction that internal friction dampens the rotational dynamics even in the limit when the polymer configuration is effectively frozen (i.e., $\tau_i \to \infty$). Fortunately, experimentally measurable dynamic properties of unfolded proteins usually only depend on internal dynamics and are unaffected by rotations.

4.5 Loop formation times

In contrast to reconfiguration times discussed in the previous section, the effect of internal friction on other timescales implicated, e.g., in protein folding is not necessarily so simple[125]. Here we explore the implications of internal friction for the simplest elementary step of folding, the formation of a contact between two sequence-distant residues of the chain.

4.5.1 Introduction to the loop formation problem.

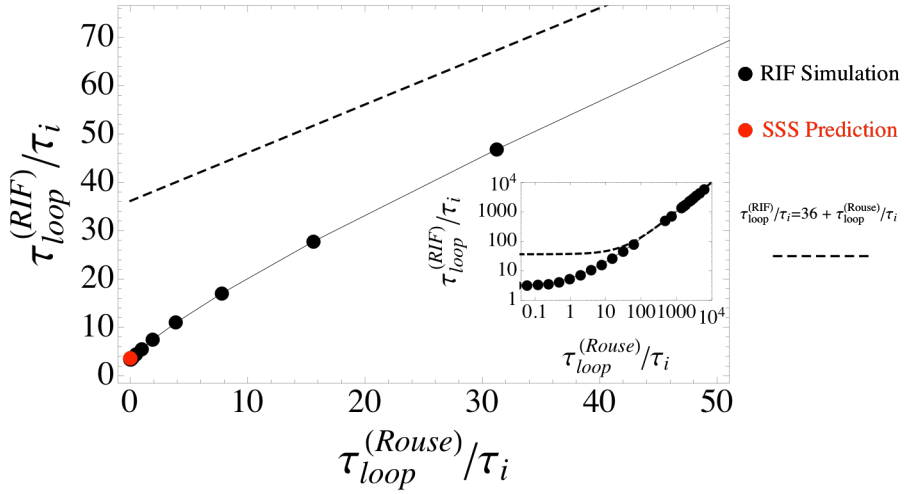
The average time τ_{loop} that it takes for the ends of a polymer to diffusively come into contact with one another and thus form a loop has been the focus of a number of theoretical[86, 107-109, 111, 112, 135-141] and experimental studies, particularly for unfolded proteins[84, 85, 87-90, 93, 95, 142, 143]. For the Rouse[109, 112] and Zimm[137] models, τ_{loop} is known to scale with chain length as $\tau_{loop}^{(Rouse)} \propto N^2$ and $\tau_{loop}^{(Zimm)} \propto N^{3/2}$ in the limit $N \to \infty$. Both models further predict τ_{loop} to be proportional to solvent viscosity, η_s. In what follows, we examine the role of internal friction on the loop formation times of RIF and ZIF. Our analysis is in part motivated by recent loop formation measurements in proteins and polypeptides[90, 91, 95, 143] as well as by the somewhat contentious issue of the effect of internal friction on the rate of protein folding[116]. While not synonymous to folding, loop formation mimics an elementary folding step[84, 86, 89, 144-146] and so our results may provide insights into the latter issue.

Two models describing the effects of internal friction on folding have been discussed in the literature. In one, internal friction has an additive effect[116, 119, 147], similar to Eqs. 4.20 and 4.21. When applied to loop formation, this implies that extrapolation of τ_{loop} to the limit of solvent absence (i.e., $\eta_s \to 0$) would result in a finite intercept that is independent of chain length and is equal to τ_i. In the other model, roughness of the energy landscape results in multiplicative renormalization of the solvent viscosity η_s [121]. In the context of loop formation, this would imply that τ_{loop} should approach zero in the

limit $\eta_s \to 0$ and exhibit the same chain length dependence as it does in the absence of internal friction. A third possibility suggested by a recent study[125] is that the internal friction effect is neither additive nor multiplicative.

In what follows we show that, consistent with ref.[125], internal and solvent friction enter into τ_{loop} in a convoluted way resulting in a nonlinear solvent viscosity dependence (Fig. 4.2), except when internal friction is weak. As a result of this nonlinearity, linear extrapolation of τ_{loop} to zero viscosity overestimates the actual loop formation time in the zero-viscosity limit.

a)



b)

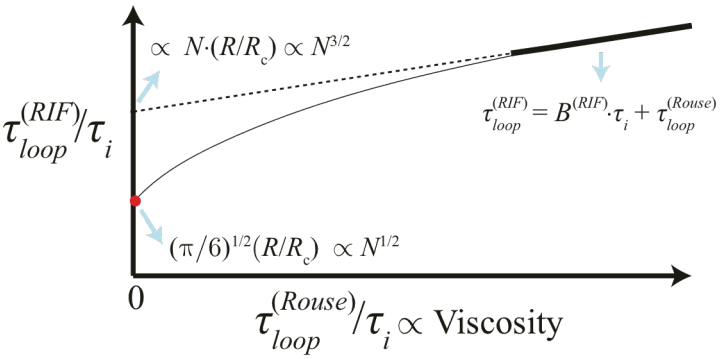


Figure 4.2: Viscosity dependence of loop formation time for RIF. a) The loop formation time plotted as a function of the dimensionless parameter, $\tau_{loop}^{(Rouse)}/\tau_i$, for an RIF chain of length $N = 40$ with a capture radius of $R_c = 1.25$. Also shown are the SSS prediction in the limit of $\tau_i \gg \tau_{Rouse}$ (red dot) given by Eq. 4.29 and the linear fit of the simulation data in the limit of $\tau_i \ll \tau_{Rouse}$ (dashed straight line). The inset contains the log-log plot of the loop formation times over the entire range of $\tau_{loop}^{(Rouse)}/\tau_i$ explored by simulation. b) This diagram highlights our key findings for the role of internal friction in the loop formation times. Here, the loop formation data given by a thin, black line is bounded by the limiting expressions in the asymptotic limits of $\tau_i \gg \tau_{Rouse}$ (red dot) and $\tau_i \ll \tau_{Rouse}$ (straight line).

4.5.2 Loop formation: Limit of high internal friction

The problem of calculating τ_{loop} is greatly simplified in the limit $\tau_i \gg \tau_{Rouse}$ or $\tau_i \gg \tau_{Zimm}$ for RIF and ZIF, respectively. In this limit, RIF (or ZIF) modes become degenerate such that Eqs. 4.8 and 4.12 can both be rewritten as

$$d\mathbf{X}_p / dt \approx -\tau_i^{-1}\mathbf{X}_p + \sigma_p \tag{4.22}$$

where σ_p is white Gaussian noise. Using Eq. 4.7, the end-to-end distance vector

$$\mathbf{R}(t) = \mathbf{r}_N(t) - \mathbf{r}_0(t) = \sum_{p=1}^{N} \mathbf{X}_p(t)\left(u_{pN} - u_{p0}\right) \tag{4.23}$$

now obeys the equation

$$d\mathbf{R} / dt = -\tau_i^{-1}\sum_{p=1}^{N} \mathbf{X}_p\left(u_{pN} - u_{p0}\right) + \sum_{p=1}^{N} \sigma_p\left(u_{pN} - u_{p0}\right),$$

or

$$d\mathbf{R} / dt = -\tau_i^{-1}\mathbf{R} + \mathbf{\Omega}(t) \tag{4.24}$$

where $\mathbf{\Omega}(t)$ is, again, a white Gaussian noise process. That is, the dynamics of the end-to-end vector is identical to that of a three-dimensional overdamped harmonic oscillator with a relaxation time of τ_i. For the purpose of calculating its end-to-end dynamics, the entire polymer can be replaced with a dumbbell consisting of two end beads attached by a single spring, a model that has been studied and discussed in refs.[109, 148].

As described in section 4.3, our definition of a "contact" is that the chain ends come within a predefined capture radius R_c. The mean first passage time τ_{loop} to attain

such a configuration can then be computed by solving the diffusion (i.e. Smoluchowski) equation for the harmonic oscillator in the 3-dimensional space, with an absorbing boundary at $|\mathbf{R}| = R_c$. Alternatively, the problem can be approached using Szabo-Schulten-Schulten (SSS) theory[107], which further reduces the dimensionality to a single degree of freedom, the absolute end-to-end distance *R*. Provided that the capture radius is much smaller than the rms end-to-end distance $\langle R^2 \rangle^{1/2}$, the SSS theory predicts the loop formation time to be given by

$$\tau_{loop} = \left(\frac{2\pi}{3} \right)^{3/2} \frac{\langle R^2 \rangle}{4\pi D R_c} \tag{4.25}$$

where *D* is the end-to-end diffusion coefficient. We estimate this coefficient using the Einstein-Stokes formula,

$$D = k_B T / \gamma \tag{4.26}$$

where γ is the effective friction coefficient of end-to-end dynamics. This friction coefficient is further related to the relaxation time τ_i and the oscillator spring coefficient *k* by

$$\tau_i = \gamma / k \tag{4.27}$$

The spring coefficient *k* can be estimated from the potential of mean force of the end-to-end vector, $\mathbf{R}$. Since the probability distribution of $\mathbf{R}$ is Gaussian, the potential of mean force is quadratic, with an effective spring coefficient of

$$k = 3k_B T / \langle R^2 \rangle \tag{4.28}$$

Substituting Eqs. 4.26, 4.27, and 4.28 into Eq. 4.25, we find

$$\tau_{loop} = \left(\frac{\pi}{6}\right)^{1/2} \tau_i \frac{\langle R^2 \rangle^{1/2}}{R_c} \tag{4.29}$$

Eq. 4.29 holds for both the RIF and ZIF models and agrees with simulation data (Fig. 4.3).

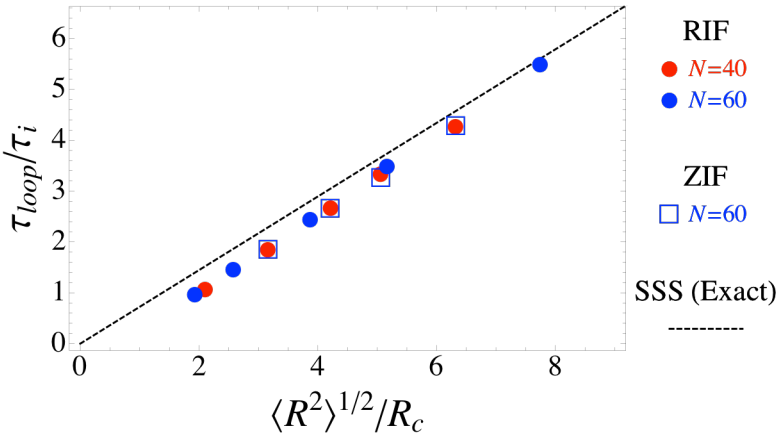


Figure 4.3: The loop formation time in high internal friction limit. Here, the end-to-end loop formation time predicted by SSS (Eq. 4.29) in the high internal friction limit is plotted as a function of the ratio $\langle R^2 \rangle^{1/2} / R_c$. Simulation data for both RIF and ZIF are plotted alongside the SSS result for comparison. All of the simulation data was found to collapse onto a single master curve, which agrees well with the SSS prediction. The agreement between simulation and SSS further improves as one approaches the limit of $R_c << \langle R^2 \rangle^{1/2}$, where the SSS formula is expected to become exact. See text for further discussion.

While it is not immediately obvious that the SSS reduction from three degrees of freedom to one would not introduce errors, τ_{loop} can also be found exactly by solving the

3-dimensional diffusion equation describing the two-bead system. The exact solution to this problem was previously given[148], albeit with a minor numerical error in the final answer. Their (corrected) solution is identical to the SSS formula (Eq. 4.29) and is shown to be exact in the limit of $R_c << \langle R^2 \rangle^{1/2}$, which is supported by the simulation data in Fig. 4.3. It should be noted that the SSS approximation is known to give incorrect scaling of τ_{loop} for the standard Rouse model[108, 112], although modifications of the theory can recover the correct N^2 dependence[110, 112]. This difficulty stems from the highly non-Markovian dynamics of the end-to-end distance[109, 149]. In contrast, the dynamics in the high internal friction limit given by Eq. 4.24 is Markovian, which is the reason why the SSS approximation recovers the exact result.

Eq. 4.29 shows that τ_{loop} exhibits a chain length dependence of $\tau_{loop} \propto N^{1/2}$ that is significantly weaker than the anticipated scaling of N^2 and $N^{3/2}$ for the Rouse and Zimm models .Yet the non-vanishing chain length dependence of τ_{loop} further shows that the scaling of τ_{loop} is different from that of the chain relaxation time, which is chain length independent and equal to τ_i when internal friction dominates.

Finally note that, although the above discussion was limited the end-to-end loops, the same arguments can be applied to the collision between an internal pair of monomers, *n* and *m*. In that case Eq. 4.29 is still valid provided that $\langle R^2 \rangle = b^2 |n-m|$. This means that the time to form an internal loop depends on the loop length $|n-m|$ but not on the

length of the tails flanking the loop. This observation is in contrast to the Rouse/Zimm cases, where the presence of long tails is known to slow down the loop closure[105, 150].

4.5.3 Loop formation: Limit of low internal friction

Consider the behavior of τ_{loop} as a function of solvent viscosity η_s. As η_s is increased, eventually the limit $\tau_{Rouse} \gg \tau_i$ is attained, where solvent friction is dominant over the internal friction (assuming that τ_i is constant). In this limit, of course, the loop formation time $\tau_{loop}^{(RIF)}$ should approach the Rouse limit of $\tau_{loop}^{(Rouse)}$, and, therefore, $\tau_{loop}^{(RIF)}(\eta_s)$ must become a linear function of solvent viscosity. When extrapolated to zero viscosity an intercept will be observed, just as in the case of nsFCS reconfiguration times, which is a signature of internal friction. This extrapolated intercept, however, is not equal to the zero-viscosity loop formation time $\tau_{loop}^{(RIF)}(\eta_s = 0)$, which is the $\tau_{Rouse} \ll \tau_i$ limit calculated in the previous section, because the dependence of $\tau_{loop}(\eta_s)$ is not a straight line (Fig. 4.2). Rather, the high-viscosity data (obtained at $\tau_{Rouse} \gg \tau_i$) extrapolated to zero viscosity gives a higher value (Fig. 4.2). Using $\tau_{loop}^{(Rouse)} \propto \eta_s$ itself as the measure of viscosity in this experiment, we write

$$\tau_{loop}^{(RIF)} = \tau_{loop}^{(Rouse)} + B^{(RIF)}\tau_i \tag{4.30a}$$

where we have surmised that the intercept $B^{(RIF)}\tau_i$ is proportional to τ_i. The dimensionless proportionality coefficient $B^{(RIF)}$ may, of course, depend on chain length

and the capture radius. It is further expedient to divide by τ_i, which results in a dimensionless form of Eq. 4.30a:

$$\tau_{loop}^{(RIF)} / \tau_i = \tau_{loop}^{(Rouse)} / \tau_i + B^{(RIF)} \tag{4.30b}$$

Written in this manner, the dimensionless loop formation time in the weak internal friction limit exhibits a linear dependence on the dimensionless parameter $\tau_{loop}^{(Rouse)} / \tau_i$ with a slope that is, by construction, 1 and with a finite intercept of $B^{(RIF)}$. If, as in Section 4.4, this contribution of internal friction were purely additive, we would expect $B^{(RIF)}$ to simply be equal to 1, but we will demonstrate next that this is not the case. Instead, this coefficient depends on both the capture radius R_c and the chain length *N*. Assuming that these dependences are power laws, we can rewrite them in a general form involving two dimensionless parameters, $\langle R^2 \rangle^{1/2} / R_c$ and *N*:

$$B^{(RIF)}(N, R_c) \propto \left(\langle R^2 \rangle^{1/2} / R_c \right)^{\delta} N^{\varepsilon} \propto N^{\varepsilon + \delta/2} \tag{4.31}$$

We have used simulations of RIF to determine both scaling exponents, δ and ε, by separately varying each parameter. In doing so, we chose to vary $\tau_{loop}^{(Rouse)} / \tau_i$ by varying τ_i rather than the solvent viscosity, which is a more convenient yet equivalent approach.

Our data (Fig. 4.4) shows that δ is close to 1. Likewise, we find that the scaling exponent ε is also close to 1 (Fig. 4.4). Thus, we find that the dimensionless intercept of Eq. 4.30b is not constant but rather exhibits a dependence on the capture radius of

$B^{(RIF)} \propto R_c^{-1}$ and a total chain length dependence of $B^{(RIF)} \propto N\left\langle R^2\right\rangle^{1/2} / R_c \propto N^{3/2}$ (where we have used $\left\langle R^2\right\rangle = Nb^2$ for the Gaussian polymer).

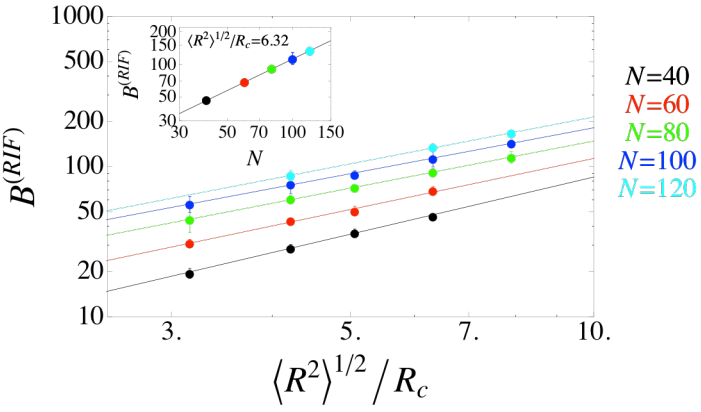


Figure 4.4. Power Law Dependence of $B^{(RIF)}$ on chain length and capture radius. The dependence of $B^{(RIF)}$ on chain length and capture radius is well described by power laws (refer to Eq. 4.31 and surrounding text for more details). As shown in the outer log-log plot, the dependence of $B^{(RIF)}$ on $\left\langle R^2\right\rangle^{1/2} / R_c$ can be fit to power law relations with scaling exponents of $\delta = 1.27 \pm 0.03$, 1.13 ± 0.06, 1.04 ± 0.01, 1.02 ± 0.02, and 1.04 ± 0.03 for $N = 40$, 60, 80, 100, and 120, respectively. Likewise, for fixed $\left\langle R^2\right\rangle^{1/2} / R_c$, the dependence of $B^{(RIF)}$ on N can be fit to power laws with scaling exponents of $\varepsilon = 1.17 \pm 0.02$, 1.04 ± 0.03, 0.99 ± 0.06, 0.966 ± 0.007, and 0.90 ± 0.03 for $\left\langle R^2\right\rangle^{1/2} / R_c = 3.16$, 4.22, 5.06, 6.32, and 7.91, respectively. The inset log-log plot shows a representative fit of the chain length dependence of $B^{(RIF)}$ when $\left\langle R^2\right\rangle^{1/2} / R_c = 6.32$. Note that our numerical values are not strictly in the limit $b \ll R_c \ll \left\langle R^2\right\rangle^{1/2}$, where Eq. 4.31 is expected to be universal, and so finite size effects may be present.

Notice that the chain length dependence of this intercept is weaker than the N^2 of the solvent controlled loop formation time but stronger than the $N^{1/2}$ dependence of the actual $\eta_s \to 0$ limit (Section 4.5.2), which is another manifestation of the nonlinearity of the viscosity dependence of the loop formation time.

Similarly to the RIF case, the loop formation time of ZIF can be expressed as

$$\tau_{loop}^{(ZIF)} / \tau_i = \tau_{loop}^{(Zimm)} / \tau_i + B^{(ZIF)} \tag{4.32}$$

where

$$B^{(ZIF)}\left(N, R_c\right) \propto \left(\left\langle R^2 \right\rangle^{1/2} / R_c\right)^{\delta} N^{\varepsilon} \tag{4.33}$$

We find from our data (Fig. 4.5) that δ is in the range $\delta \approx 1.15 - 1.38$ for $N = 60 - 120$ and ε is in the range $\varepsilon \approx 0.63 - 0.92$ for $\left\langle R^2 \right\rangle^{1/2} / R_c \approx 3.16 - 7.91$. The variation in these indices is presumably due to finite-size effects. Indeed, the end-to-end loop formation times of the Zimm chain was found to exhibit a chain length dependence of $\propto N^{1.3}$ (data not shown) instead of the theoretically predicted $\propto N^{3/2}$ universal law expected in the limit of $N \to \infty$ [137]. Notwithstanding deviations from universality, the overall chain length dependence $B^{(ZIF)} \propto N^{\varepsilon+\delta/2} = N^{1.2} - N^{1.6}$ is roughly the same as in the RIF case, while the capture radius dependence is $B^{(ZIF)} \propto R_c^{-1.15} - R_c^{-1.38}$.

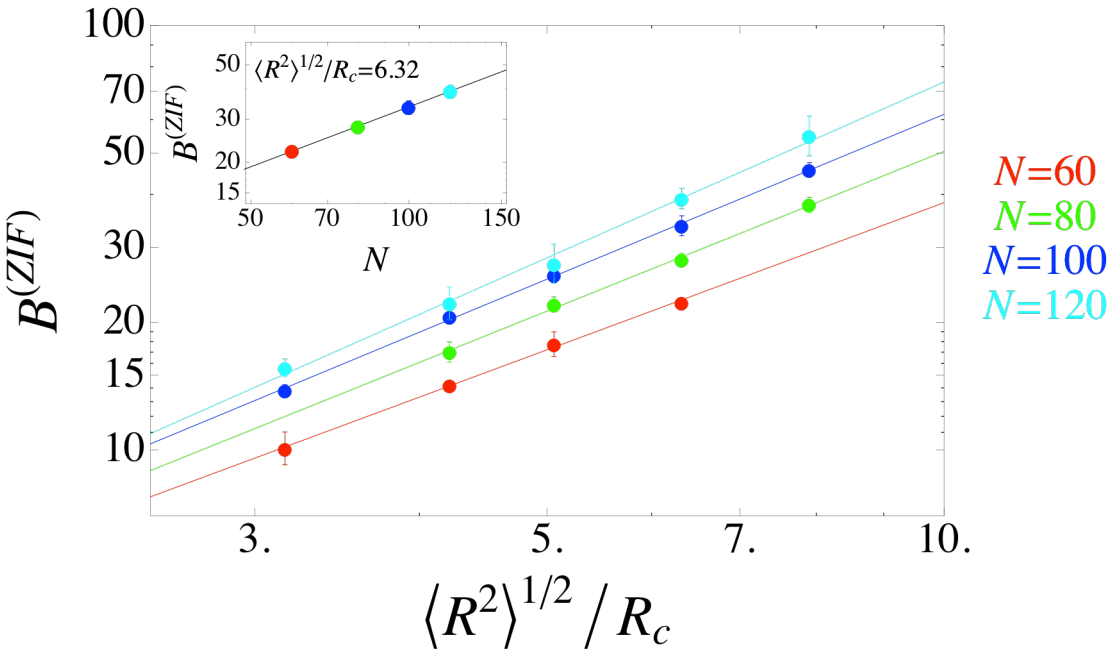


Figure 4.5: Power Law Dependence of $B^{(ZIF)}$ on chain length and capture radius. By systematically varying chain length and capture radius for ZIF, we found that $B^{(ZIF)}$ is well described by the power law dependences shown by Eq. 4.33 (see text for details). In the outer log-log plot, the dependence of $B^{(ZIF)}$ on $\langle R^2\rangle^{1/2}/R_c$ can be fit to power laws with scaling exponents of $\delta = 1.15 \pm 0.03$, 1.25 ± 0.04, 1.29 ± 0.02, and 1.38 ± 0.05 for $N = 60$, 80, 100, and 120, respectively. When $\langle R^2\rangle^{1/2}/R_c$ is fixed, $B^{(ZIF)}$ is found to scale as a power law of N with the scaling exponents $\varepsilon = 0.629 \pm 0.009$, 0.66 ± 0.05, 0.65 ± 0.06, 0.813 ± 0.004, and 0.92 ± 0.05 for $\langle R^2\rangle^{1/2}/R_c = 3.16$, 4.22, 5.06, 6.32, and 7.91, respectively. A representative fit of $B^{(ZIF)}$ when $\langle R^2\rangle^{1/2}/R_c = 6.32$ is shown in the inset log-log plot.

4.5.4 Loop formation: Summary of viscosity and internal friction dependence

The viscosity and internal friction dependence of the loop formation time is summarized in Fig. 4.2 for the RIF case. At high viscosity, τ_{loop} exhibits a linear viscosity dependence. When extrapolated from high viscosity to zero viscosity, this dependence exhibits an intercept $B^{(RIF)}\tau_i$, which reflects internal friction. This intercept, however, is much larger than τ_i (for $N \gg 1$) in contrast to the additive internal friction model. At the same time, the existence of an intercept and the non-vanishing loop formation time in the zero-viscosity limit also contradicts the multiplicative model, which predicts τ_{loop} to approach zero as $\eta_s \to 0$.

In contrast to the linear growth of τ_{loop} with viscosity found at high viscosity, the actual viscosity dependence of $\tau_{loop}(\eta_s)$ is highly nonlinear, resulting in a zero-viscosity limit, $\tau_{loop}(0)$, that is much smaller than the intercept $B^{(RIF)}\tau_i$ obtained via extrapolation of high-viscosity data. Instead, $\tau_{loop}(0)$ displays an even weaker chain length dependence of $N^{1/2}$ and is inversely proportional to the capture radius, a result rigorously proven in Section 4.5.2.

Internal friction is just one possible cause of nonlinearity of the curve $\tau_{loop}(\eta_s)$. Competition between quenching kinetics and intramolecular diffusion, for example, may also lead to a nonlinear viscosity dependence[151], which is well approximated by a power law, $\tau_{loop}(\eta_s) \propto \eta_s^{\alpha}, \alpha < 1$, that has a shape qualitatively similar to that in Fig. 4.2.

Although the dependence $\tau_{loop}(\eta_s)$ presented here differs from a power law in that it does not approach zero as $\eta_s \to 0$ and it becomes linear as $\eta_s \to \infty$, it may be difficult to distinguish between the two cases in practice, given a typically limited range of viscosities accessible experimentally. Thus any experimental observations of curvature of the dependence $\tau_{loop}(\eta_s)$ should be regarded with caution.

4.5.5 Comparison with experiments

Do experimental measurements of loop formation times in unfolded polypeptides bear signature of internal friction? As internal friction effects tend to be more pronounced for short chains, we have looked for signatures of internal friction in the data reported by the Kiefhaber group[90, 91], who examined unfolded poly(Gly-Ser) peptides consisting of short, 3 to 14 repeat units at different concentrations of denaturant. Based on the results of ref.[94], we further expect that internal friction effects, if any, should be more pronounced at low denaturant. However, decreasing the denaturant concentration was also found to result in a decrease in the dimensions of the polypeptide, an effect that reduces the loop formation time by enhancing the spatial proximity of the loop ends. To model the latter effect, we have subjected each monomer of the RIF chain to an attractive harmonic potential, as discussed in the Introduction and further detailed in the Appendix. The strength of this potential was varied to reproduce the experimentally observed dimensions of the unfolded peptides[91].

With this compaction effect accounted for, we find remarkable agreement

between our simulated data and the experimental data, provided that internal friction is assumed to be completely absent (Fig. 4.6).

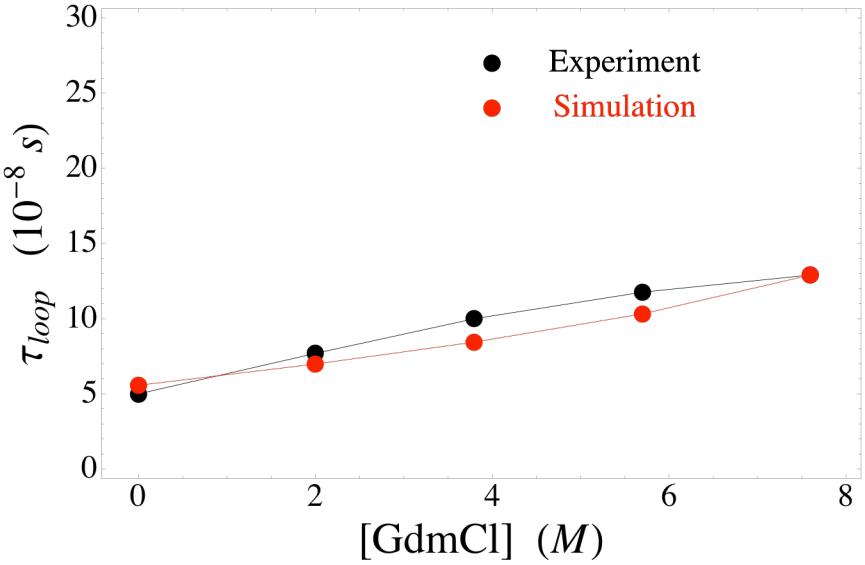


Figure 4.6: Comparison of simulated and experimental loop formation times. Experimental end-to-end contact formation times of a $(Gly\text{-}Ser)_{14}$ peptide, as a function of the denaturant concentration[90], agrees well with the CRIF model (see Appendix for details) assuming zero internal friction. The model was parameterized to reproduce the experimentally measured end-to-end distance of the peptide at each denaturant concentration[91]. Here, the simulated polymer had a chain length of $N = 6$ and the Kuhn length was chosen to be equal to $b = 19$ Å (i.e., roughly 5 amino acids). The value of the capture radius of $R_c \approx 3$ Å was chosen to match the experimental data at 7.6M GdmCl.

That is, the speedup in the loop formation times with decreasing denaturant concentration occurs entirely because of the enhanced spatial proximity of the loop-forming groups afforded by chain compaction, consistent with the argument made in refs.[90, 91]. While we

would expect internal friction to be more prominent at lower denaturant concentrations, thereby opposing the speedup caused by chain compaction, we did not observe such a trend in our comparison, suggesting that internal friction effects are negligible for this particular experimental system. This conclusion is not a surprise considering the reported near-direct proportionality between loop formation times and solvent viscosity (i.e., no intercept)[90].

We do not expect, however, this finding to be general. It is plausible that foldable polypeptide sequences (i.e., proteins) would exhibit more pronounced internal friction that could be observed in loop formation experiments. Indeed, recent work from the Lapidus group[95] suggests that the internal friction effect on loop formation times in unfolded proteins can be quite significant under conditions close to native.

4.6 Concluding remarks

Motivated by the recent success of RIF in describing FRET-derived relaxation timescales of unfolded proteins under near native conditions[94], here we have extended this model to include hydrodynamic effects and explored its predictions for other types of experiments, particularly those probing the rates at which sequence-distant segments of a polypeptide chain form a close contact. We find that the manner in which internal friction manifests itself in the latter type of experiments is generally neither additive nor multiplicative, in contrast to the predictions of simpler, phenomenological models. Thus our theory offers a nontrivial experimental test of ZIF/RIF through a systematic study of

loop closure rates in unfolded proteins, of the type reported, e.g., in refs. [88, 90, 95, 152-157]. Such experimental validation is particularly important since the definitive elucidation of internal friction and its microscopic origins through detailed molecular simulations is generally hampered by the high computational cost of such studies. For example, the typical timescales of internal friction, τ_i, observed in ref. [94], which are in a 10-100 ns range, challenge atomistic simulations of those systems since simulation times that are orders of magnitude longer than τ_i would be required to generate converged results and test the specific predictions of RIF or ZIF and since enhanced sampling methods such as replica exchange are unsuitable for computing dynamical properties.

Given the abundance of intrinsically disordered proteins, protein dynamics in the unfolded state is of significant interest in its own right. Yet much of the past investigation into internal friction in proteins focused on its implications for folding. While addressing this issue is beyond the scope of the present work, it should be noted that the formation of a single contact may be viewed as the simplest model (or at least a caricature) of folding. For example, a model of this type, called Generalized Rouse Model, has been used to study various aspects of biomolecular folding/unfolding dynamics[158, 159]. It is then plausible that some of the findings reported here may shed light on the highly discussed role of internal friction in folding. In particular, our results suggest that faster folding processes would be impacted more strongly by internal friction, which may explain the lack of detectable internal friction effects in earlier folding studies[160]. Of course, the same prediction is made by the additive internal friction model[116, 119, 147], where the effect of the

internal friction on the folding time τ_f amounts to an additive constant, $\tau_f \to \tau_f + \tau_i$. Thus for slow folding, where $\tau_f \gg \tau_i$, the effect of internal friction is predicted to become negligible. Our results for loop closure, however, show that this additive model is inadequate. That is, if one extrapolates the loop closure time to zero solvent viscosity, the resulting intercept is not simply equal to the internal friction time τ_i but rather itself depends on the entropic cost of forming the loop. For example, increasing the polymer length to slow down loop closure will also magnify the internal friction effect. However the internal friction contribution into the contact formation time tends to have weaker chain length dependence than that of the solvent friction and so the relative effect of internal friction still vanishes in the $N \to \infty$ limit.

Although structural biases and sequence effects are arguably important in the dynamics of unfolded proteins, particularly under conditions close to native, this and earlier[94, 105] work suggest that such effects, at least on the average, may already be captured by polymer theories. In particular, introduction of a non-specific phenomenological internal friction time constant and a simple compressive potential seem to capture the experimentally observable trends in both the chain reconfiguration dynamics and statistics. It should be emphasized, however, that molecular understanding of these phenomena remains limited and further insights will likely require extensive and challenging atomistic simulations.

4.7 Appendix. Compacted Rouse with internal friction (CRIF)

When the amount of denaturant is decreased, the spatial dimensions of an unfolded protein are expected to decrease. While this effect can be captured by coil-globule transition theories[104], those theories deal with statistical properties rather than the temporal evolution of the polymer. To account for the collapse of the chain without sacrificing the simplicity of the RIF model we adopt the approach described in ref. [94] to augment Eq. 4.4 with a potential that acts to pull each monomer towards the coordinate origin:

$$V_c = \frac{1}{2} k_c \sum_{n=0}^{N} r_n^2 \tag{4.34}$$

where k_c is the spring constant of this potential. We refer to this model as a "compacted Rouse model with internal friction", or CRIF. Of course, this potential breaks translational symmetry and so CRIF would be inadequate as a description of translational motion of the protein, but it remains a physically sensible description of its internal dynamics. As previously discussed[94], a potential of this form does not change the eigenmodes of the Rouse model and only modifies the equation of motion for each mode (cf. Eq. 4.8) by shifting the associated spring constant to $k_p \to \kappa\lambda_p + k_c$. Various statistical and dynamical properties of the CRIF chain, such as the spectrum of reconfiguration times, can then be obtained in much the same way as for RIF (cf. Eqs. 4.14-4.19). In particular, the rms distance between a pair of monomers, *n* and *m*, can be obtained from Eq. 4.7:

$$\left\langle \left(\mathbf{r}_n - \mathbf{r}_m\right)^2 \right\rangle_{CRIF}^{1/2} = \left(3k_B T \sum_{p=1}^{N} \left(k_c + k_0 \lambda_p\right)^{-1} \left(u_{pn} - u_{pm}\right)^2 \right)^{1/2} \qquad (4.35)$$

where the equipartition theorem, stating that the mean potential energy stored in a three-dimensional harmonic oscillator equals $k_p \left\langle \mathbf{X}_p^2 \right\rangle / 2 = 3k_B T / 2$, has been used. CRIF was found to correctly reproduce the collapse-induced changes in the chain statistics computed for more realistic polymers undergoing the coil-globule transition [94]. In this respect, CRIF is a more realistic approach to modeling chain collapse than simply reducing the Kuhn length *b* or, equivalently, increasing the bond spring constant k_0. Indeed, in the latter case the Gaussian random-coil statistics would be preserved so that $\left\langle (\mathbf{r}_n - \mathbf{r}_m)^2 \right\rangle^{1/2} = b|n-m|^{1/2}$, a result inconsistent with simulations as well as with physical intuition [94].

In our comparisons with experimental data we assume that the chain is a Gaussian random coil (i.e. $k_c = 0$) at high denaturant concentrations. We then use k_c as an adjustable parameter at lower denaturant concentrations such that the dimensions of the CRIF chain given by Eq. 4.35 matches experimentally observed values of the rms end-to-end distance.

References

1. A. E. Cardenas and R. Elber, Proteins 51 (2), 245-257 (2003).
2. X. W. Fang, P. Thiyagarajan, T. R. Sosnick and T. Pan, P Natl Acad Sci USA 99 (13), 8518-8523 (2002).
3. D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan and W. Wriggers, Science 330 (6002), 341-346 (2010).
4. P. Majek and R. Elber, J Chem Theory Comput 6 (6), 1805-1817 (2010).
5. A. M. A. West, R. Elber and D. Shalloway, J Chem Phys 126 (14), 145104 (2007).
6. S. M. Kreuzer, R. Elber and T. J. Moon, J Phys Chem B 116 (29), 8662-8691 (2012).
7. G. S. Jas, W. A. Hegefeld, P. Majek, K. Kuczera and R. Elber, J Phys Chem B 116 (23), 6598-6610 (2012).
8. A. E. Cardenas, G. S. Jas, K. Y. DeLeon, W. A. Hegefeld, K. Kuczera and R. Elber, J Phys Chem B 116 (9), 2739-2750 (2012).
9. S. Kirmizialtin, V. Nguyen, K. A. Johnson and R. Elber, Structure 20 (4), 618-627 (2012).
10. H. S. Chung, J. M. Louis and W. A. Eaton, P Natl Acad Sci USA 106 (29), 11837-11844 (2009).
11. H. S. Chung, K. McHale, J. M. Louis and W. A. Eaton, Science 335 (6071), 981-984 (2012).
12. T. H. Lee, L. J. Lapidus, W. Zhao, K. J. Travers, D. Herschlag and S. Chu, Biophysical Journal 92 (9), 3275-3283 (2007).
13. K. Neupane, D. B. Ritchie, H. Yu, D. A. N. Foster, F. Wang and M. T. Woodside, Physical Review Letters 109 (6), 068102 (2012).

14. X. L. Hao Yu, Krishna Neupane, Amar Nath Gupta, Angela M. Brigley, Allison Solanki, Iveta Sosova, and Michael T. Woodside, Proc. Natl Acad. Sci. USA 109 (14), 5283-5288 (2012).
15. E. R. Bazúa and M. C. Williams, The Journal of Chemical Physics 59 (6), 2858-2868 (1973).
16. P.-G. d. Gennes, *Scaling Concepts in Polymer Physics*. (Cornell University Press, Ithaca, 1979).
17. C. W. Manke and M. C. Williams, Macromolecules 18, 2045-2051 (1985).
18. J. J. Portman, S. Takada and P. G. Wolynes, The Journal of Chemical Physics 114 (11), 5082-5096 (2001).
19. B. S. Khatri and T. C. B. McLeish, Macromolecules 40 (18), 6770-6777 (2007).
20. A. K. Faradjian and R. Elber, J Chem Phys 120 (23), 10880-10889 (2004).
21. E. Vanden-Eijnden, M. Venturoli, G. Ciccotti and R. Elber, J Chem Phys 129 (17), 174102 (2008).
22. E. Vanden-Eijnden and M. Venturoli, J Chem Phys 130 (19), 194101 (2009).
23. A. T. Hawk and D. E. Makarov, J Chem Phys 135 (22) (2011).
24. R. Elber and A. West, P Natl Acad Sci USA 107 (11), 5001-5005 (2010).
25. P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, Annu Rev Phys Chem 53, 291-318 (2002).
26. D. Moroni, P. G. Bolhuis and T. S. van Erp, J Chem Phys 120 (9), 4055-4065 (2004).
27. D. Moroni, T. S. van Erp and P. G. Bolhuis, Physica A 340 (1-3), 395-401 (2004).
28. R. J. Allen, D. Frenkel and P. R. ten Wolde, J Chem Phys 124 (19), 024102 (2006).
29. R. J. Allen, D. Frenkel and P. R. ten Wolde, J Chem Phys 124 (2), 024102 (2006).
30. T. S. van Erp, D. Moroni and P. G. Bolhuis, J Chem Phys 118 (17), 7762-7774 (2003).
31. N. S. Hinrichs and V. S. Pande, J Chem Phys 126 (24), 244101-244111 (2007).

32. N. V. Buchete and G. Hummer, J Phys Chem B 112 (19), 6057-6069 (2008).
33. M. Sarich, F. Noe and C. Schutte, Multiscale Model Sim 8 (4), 1154-1177 (2010).
34. J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte and F. Noe, J Chem Phys 134 (17), - (2011).
35. C. Schutte, F. Noe, J. F. Lu, M. Sarich and E. Vanden-Eijnden, J Chem Phys 134 (20), - (2011).
36. L. Huang and D. E. Makarov, J Chem Phys 128 (11) (2008).
37. W. W. Zheng, M. A. Rohrdanz, M. Maggioni and C. Clementi, J Chem Phys 134 (14) (2011).
38. A. T. Hawk, S. Konda and D. E. Makarov, (2012).
39. A. N. Shiryaev, *Probability*. (Springer-Verlag, New York, 1996).
40. J. A. Hartigan, *Clustering Algorithms*. (Wiley, New York, 1975).
41. D. Mackay, *Information Theory, inference and Learning Algorithms*. (Cambridge university Press, 2003).
42. P. S. Bradley, K. P. Bennet and A. Demiriz, Technical Report, 2000.
43. G. Hummer, J Chem Phys 120 (2), 516-523 (2004).
44. S. Kirmizialtin and R. Elber, J. Phys. Chem. A 115 (Copyright (C) 2011 American Chemical Society (ACS). All Rights Reserved.), 6137-6148 (2011).
45. T. S. van Erp and P. G. Bolhuis, J Comput Phys 205 (1), 157-181 (2005).
46. C. Dellago, P. G. Bolhuis and P. L. Geissler, Adv Chem Phys 123, 1-78 (2002).
47. S. Kirmizialtin and R. Elber, J. Phys. Chem. A 115, 6137-6148 (2011).
48. A. Meller, Journal of Physics: Condensed Matter 15 (17) (2003).
49. J. Mathe, H. Visram, V. Viasnoff, Y. Rabin and A. Meller, Biophys J 87 (5), 3205-3212 (2004).
50. A. Meller and D. Branton, ELECTROPHORESIS 23 (16), 2583-2591 (2002).
51. A. Meller, L. Nivon and D. Branton, Physical Review Letters 86 (15), 3435-3438 (2001).

52. T. Ambjornsson, S. P. Apell, Z. Konkoli, E. A. Di Marzio and J. J. Kasianowicz, The Journal of Chemical Physics 117 (8), 4063-4073 (2002).
53. J. J. Kasianowicz, E. Brandin, D. Branton and D. W. Deamer, Proc. Natl Acad. Sci. USA 93 (24), 13770-13773 (1996).
54. A. J. Storm, J. H. Chen, H. W. Zandbergen and C. Dekker, Physical Review E 71 (5), 051903 (2005).
55. Mohammad, S. Prakash, A. Matouschek and L. Movileanu, Journal of the American Chemical Society 130 (12), 4081-4088 (2008).
56. M. Mohammad and L. Movileanu, European Biophysics Journal 37 (6), 913-925 (2008).
57. L. Movileanu, Trends in Biotechnology 27 (6), 333-341 (2009).
58. L. Movileanu, Soft Matter 4 (5), 925-931 (2008).
59. R. Stefureac, Y.-t. Long, H.-B. Kraatz, P. Howard and J. S. Lee, Biochemistry 45 (30), 9172-9179 (2006).
60. D. M. Zuckerman and T. B. Woolf, J Chem Phys 116 (6), 2586-2591 (2002).
61. A. M. Berezhkovskii, G. Hummer and S. M. Bezrukov, Physical Review Letters 97 (2) (2006).
62. S. Chaudhury and D. E. Makarov, J Chem Phys 133 (3) (2010).
63. B. W. Zhang, D. Jasnow and D. M. Zuckerman, J Chem Phys 126 (7) (2007).
64. M. Andrey, Journal of Physics: Condensed Matter 23 (10), 103101 (2011).
65. J. K. Wolterink, G. T. Barkema and D. Panja, Physical Review Letters 96 (20), 208301 (2006).
66. J. Chuang, Y. Kantor and M. Kardar, Physical Review E 65 (1), 011802 (2001).
67. D. Wei, W. Yang, X. Jin and Q. Liao, The Journal of Chemical Physics 126 (20), 204901-204908 (2007).
68. S. Guillouzic and G. W. Slater, Physics Letters A 359 (4), 261-264 (2006).
69. P. Debabrata, T. B. Gerard and C. B. Robin, Journal of Physics: Condensed Matter 19, 432202 (2007).

70. K. Luo, S. T. T. Ollila, I. Huopaniemi, T. Ala-Nissila, P. Pomorski, M. Karttunen, S.-C. Ying and A. Bhattacharya, Physical Review E 78 (5), 050901 (2008).

71. V. V. Lehtola, R. P. Linna and K. Kaski, Physical Review E 81 (3), 031803 (2010).

72. S. V. Malinin and V. Y. Chernyak, The Journal of Chemical Physics 132 (1), 014504-014511 (2010).

73. D. Panja, Journal of Statistical Mechanics: Theory and Experiment 2010 (06), P06011 (2010).

74. D. Panja and G. T. Barkema, The Journal of Chemical Physics 132 (1), 014902 (2010).

75. H. W. de Haan and G. W. Slater, The Journal of Chemical Physics 136 (15), 154903-154912 (2012).

76. C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. (Springer-Verlag, Berlin, 1983).

77. J. L. A. Dubbeldam, A. Milchev, V. G. Rostiashvili and T. A. Vilgis, Europhysics Letters 79 (1) (2007).

78. S. S. M. Konda, A. T. Hawk and D. E. Makarov, In preparation.

79. S. Kirmizialtin, L. Huang and D. E. Makarov, Phys Status Solidi B 243 (9), 2038-2047 (2006).

80. D. E. Makarov, Accounts Chem Res 42 (2), 281-289 (2009).

81. S. Matysiak, A. Montesi, M. Pasquali, A. B. Kolomeisky and C. Clementi, Physical Review Letters 96 (11) (2006).

82. A. T. Hawk, (2012).

83. J. Kubelka, J. Hofrichter and W. A. Eaton, Current Opinion in Structural Biology 14 (1), 76-88 (2004).

84. O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber, Proceedings of the National Academy of Sciences 96 (17), 9597-9601 (1999).

85. L. J. Lapidus, W. A. Eaton and J. Hofrichter, Proceedings of the National Academy of Sciences 97 (13), 7220-7225 (2000).
86. D. Thirumalai, The Journal of Physical Chemistry B 103 (4), 608-610 (1999).
87. L. J. Lapidus, P. J. Steinbach, W. A. Eaton, A. Szabo and J. Hofrichter, The Journal of Physical Chemistry B 106 (44), 11628-11640 (2002).
88. B. Fierz and T. Kiefhaber, J Am Chem Soc 129, 672-679 (2007).
89. S. J. Hagen, J. Hofrichter, A. Szabo and W. A. Eaton, Proceedings of the National Academy of Sciences 93 (21), 11615-11617 (1996).
90. A. Moglich, F. Krieger and T. Kiefhaber, Journal of Molecular Biology 345 (1), 153-162 (2005).
91. A. Moglich, K. Joder and T. Kiefhaber, Proceedings of the National Academy of Sciences 103 (33), 12394-12399 (2006).
92. B. Schuler and W. A. Eaton, Current Opinion in Structural Biology 18, 16-26 (2008).
93. A. Soranno, R. Longhi, T. Bellini and M. Buscaglia, Biophys J 96 (4), 1515-1528 (2009).
94. A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Muller-Spath, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov and B. Schuler, Proceedings of the National Academy of Sciences (2012).
95. S. A. Waldauer, O. Bakajin and L. J. Lapidus, Proceedings of the National Academy of Sciences 107 (31), 13713-13717 (2010).
96. D. Nettels, I. V. Gopich, A. Hoffmann and B. Schuler, Proceedings of the National Academy of Sciences 104 (8), 2655-2660 (2007).
97. D. Nettels, A. Hoffmann and B. Schuler, J Phys Chem B 112 (19), 6137-6146 (2008).
98. M. Andrec, A. K. Felts, E. Gallicchio and R. M. Levy, Proc Natl Acad Sci U S A 102 (19), 6801-6806 (2005).

99. D. A. Potoyan and G. A. Papoian, Journal of the American Chemical Society 133 (19), 7405-7415 (2011).
100. I. S. Millet, S. Doniach and K. W. Plaxco, Adv. Protein Chem. 62, 241-262 (2002).
101. E. R. McCarney, J. H. Werner, S. L. Bernstein, I. Ruczinski, D. E. Makarov, P. M. Goodwin and K. W. Plaxco, J Mol Biol 352 (3), 672-682 (2005).
102. Z. Wang, K. W. Plaxco and D. E. Makarov, Biopolymers 86 (4), 321-328 (2007).
103. E. P. O'Brien, G. Morrison, B. R. Brooks and D. Thirumalai, The Journal of Chemical Physics 130 (12), 124903-124910 (2009).
104. G. Ziv, D. Thirumalai and G. Haran, Physical Chemistry Chemical Physics 11 (1), 83-93 (2009).
105. R. R. Cheng, T. Uzawa, K. W. Plaxco and D. E. Makarov, Biophysical journal 99 (12), 3959-3968 (2010).
106. D. E. Makarov, The Journal of Chemical Physics 132, 035104 (2010).
107. A. Szabo, K. Schulten and Z. Schulten, J. Chem. Phys. 72, 4350-4357 (1980).
108. R. W. Pastor, R. Zwanzig and A. Szabo, The Journal of Chemical Physics 105 (9), 3878-3882 (1996).
109. M. Doi, Chemical Physics 9 (3), 455-466 (1975).
110. R. R. Cheng and D. E. Makarov, The Journal of Chemical Physics 134 (8), 085104 (2011).
111. J. J. Portman, J. Chem. Phys. 118, 2381-2391 (2003).
112. N. M. Toan, G. Morrison, C. Hyeon and D. Thirumalai, The Journal of Physical Chemistry B 112 (19), 6094-6106 (2008).
113. T. Cellmer, E. R. Henry, J. Hofrichter and W. A. Eaton, Proceedings of the National Academy of Sciences 105 (47), 18320-18325 (2008).
114. A. Ansari, J. Chem. Phys. 110, 1774-1780 (1999).
115. A. Ansari, C. M. Jones, E. R. Henry, J. Hofrichter and W. A. Eaton, Science 256 (5065), 1796-1798 (1992).

116. S. J. Hagen, Current Protein and Peptide Science 11 (5), 1-11 (2010).
117. B. G. Wensley, S. Batey, F. A. C. Bone, Z. M. Chan, N. R. Tumelty, A. Steward, L. G. Kwa, A. Borgia and J. Clarke, Nature 463 (7281), 685-688 (2010).
118. G. S. Jas, W. A. Eaton and J. Hofrichter, The Journal of Physical Chemistry B 105 (1), 261-272 (2000).
119. L. Qiu and S. J. Hagen, Journal of the American Chemical Society 126 (11), 3398-3399 (2004).
120. D. E. Sagnella, J. E. Straub and D. Thirumalai, The Journal of Chemical Physics 113 (17), 7702-7711 (2000).
121. R. Zwanzig, Proceedings of the National Academy of Sciences 85 (7), 2029-2030 (1988).
122. R. B. Best and G. Hummer, Physical Review Letters 96 (22), 228104 (2006).
123. A. Erbas, D. Horinek and R. R. Netz, Journal of the American Chemical Society 134 (1), 623-630 (2011).
124. B. S. Khatri, M. Kawakami, K. Byrne, D. A. Smith and T. C. B. McLeish, Biophysical journal 92 (6), 1825-1835 (2007).
125. J. C. F. Schulz, L. Schmidt, R. B. Best, J. Dzubiella and R. R. Netz, Journal of the American Chemical Society 134 (14), 6273-6279 (2012).
126. M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics*. (Oxford University Press, 1986).
127. A. Hoffmann, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D. A. Horsley, R. Seckler, O. Bakajin and B. Schuler, Proc Natl Acad Sci U S A 104 (1), 105-110 (2007).
128. S. Muller-Spath, A. Soranno, V. Hirschfeld, H. Hofmann, S. Ruegger, L. Reymond, D. Nettels and B. Schuler, Proc Natl Acad Sci U S A 107 (33), 14609-14614 (2010).

129. D. Nettels, S. Muller-Spath, F. Kuster, H. Hofmann, D. Haenni, S. Ruegger, L. Reymond, A. Hoffmann, J. Kubelka, B. Heinz, K. Gast, R. B. Best and B. Schuler, Proc Natl Acad Sci U S A (2009).
130. G. Ziv and G. Haran, J Am Chem Soc 131 (8), 2942-2947 (2009).
131. Z. S. Wang and D. E. Makarov, J Phys Chem B 107, 5617-5622 (2003).
132. C. L. Ting and D. E. Makarov, J Chem Phys 128 (11), 115102 (2008).
133. I. V. Gopich, D. Nettels, B. Schuler and A. Szabo, The Journal of Chemical Physics 131 (9) (2009).
134. B. G. Wensley, L. G. Kwa, S. L. Shammas, J. M. Rogers, S. Browning, Z. Yang and J. Clarke, Proceedings of the National Academy of Sciences 109 (44), 17795-17799 (2012).
135. G. Wilemski and M. Fixman, J. Chem. Phys. 60, 866-877 (1974).
136. G. Wilemski and M. Fixman, J. Chem. Phys. 60, 878-890 (1974).
137. B. Friedman and B. O'Shaughnessy, Physical Review A 40 (10), 5950-5959 (1989).
138. A. Perico and M. Beggiato, Macromolecules 23, 797-803 (1990).
139. M. Ortiz-Repiso, J. J. Freire and A. Rey, Macromolecules (31), 8356-8362 (1998).
140. M. Ortiz-Repiso and A. Rey, Macromolecules (31), 8363-8369 (1998).
141. A. Podtelezhnikov and A. Vologodskii, Macromolecules 30 (21), 6668-6673 (1997).
142. T. Uzawa, R. R. Cheng, K. J. Cash, D. E. Makarov and K. W. Plaxco, Biophysical journal 97 (1), 205-210 (2009).
143. M. Buscaglia, L. J. Lapidus, W. A. Eaton and J. Hofrichter, Biophys J 91, 276-288 (2006).
144. D. E. Makarov and K. W. Plaxco, Protein Sci. 12, 17 (2003).
145. D. E. Makarov, C. Keller, K. W. Plaxco and H. Metiu, Proc. Natl. Acad. Sci. U.S.A. 99, 3535 (2002).

146. M. Buscaglia, J. Kubelka, W. A. Eaton and J. Hofrichter, J Mol Biol 347 (3), 657-664 (2005).
147. S. J. Hagen, L. Qiu and S. A. Pabit, J. Phys. Condens. Matter 17, S1503-S1514 (2005).
148. S. Sunagawa and M. Doi, Polymer Journal 7 (6), 604-612 (1975).
149. I. M. Sokolov, Phys Rev Lett 90 (8), 080601 (2003).
150. B. Friedman and B. O'Shaughnessy, Macromolecules 26 (18), 4888-4898 (1993).
151. R. R. Cheng, T. Uzawa, K. W. Plaxco and D. E. Makarov, The Journal of Physical Chemistry B 113 (42), 14026-14034 (2009).
152. I. Daidone, H. Neuweiler, S. Doose, M. Sauer and J. C. Smith, PLoS Comput Biol 6 (1), e1000645 (2010).
153. H. Neuweiler, A. Schulz, M. Bohmer, J. Enderlein and M. Sauer, Journal of the American Chemical Society 125 (18), 5324-5330 (2003).
154. B. Fierz, A. Reiner and T. Kiefhaber, Proc Natl Acad Sci U S A 106 (4), 1057-1062 (2009).
155. A. Reiner, P. Henklein and T. Kiefhaber, Proc Natl Acad Sci U S A 107 (11), 4955-4960 (2010).
156. B. Ahmad, Y. Chen and L. J. Lapidus, Proceedings of the National Academy of Sciences of the United States of America 109 (7), 2336-2341 (2012).
157. V. A. Voelz, M. Jager, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss and V. S. Pande, Journal of the American Chemical Society 134 (30), 12565-12577 (2012).
158. V. Barsegov, G. Morrison and D. Thirumalai, Physical review letters 100 (24), 248102 (2008).
159. C. Hyeon, G. Morrison and D. Thirumalai, Proceedings of the National Academy of Sciences of the United States of America 105 (28), 9604-9609 (2008).
160. K. W. Plaxco and D. Baker, Proceedings of the National Academy of Sciences of the United States of America 95 (23), 13591-13596 (1998).